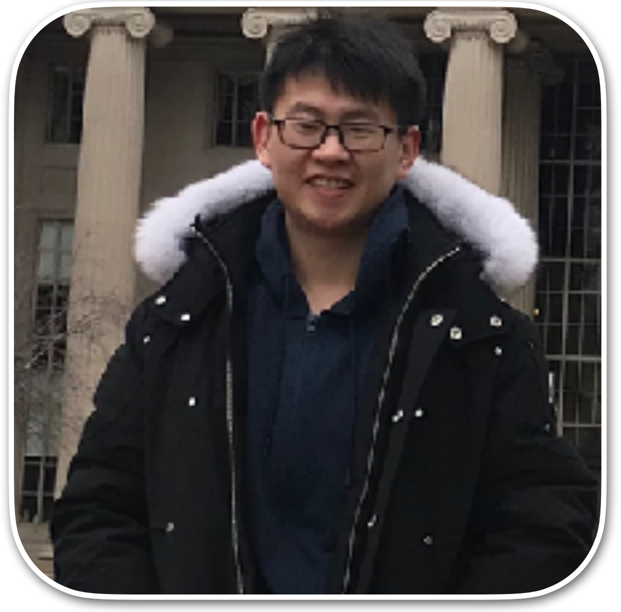
Theoretically Grounded Framework for LLM Watermarking: A Distribution-Adaptive Approach

Haiyun He
Postdoc @ Center for Applied Mathematics, Cornell University

ITA 2025

Yepeng Liu
Univ. of Florida



Prof. Ziqiao Wang
Tongji Univ.

Prof. Yongyi Mao
Univ. of Ottawa

Prof. Yuheng Bu
Univ. of Florid

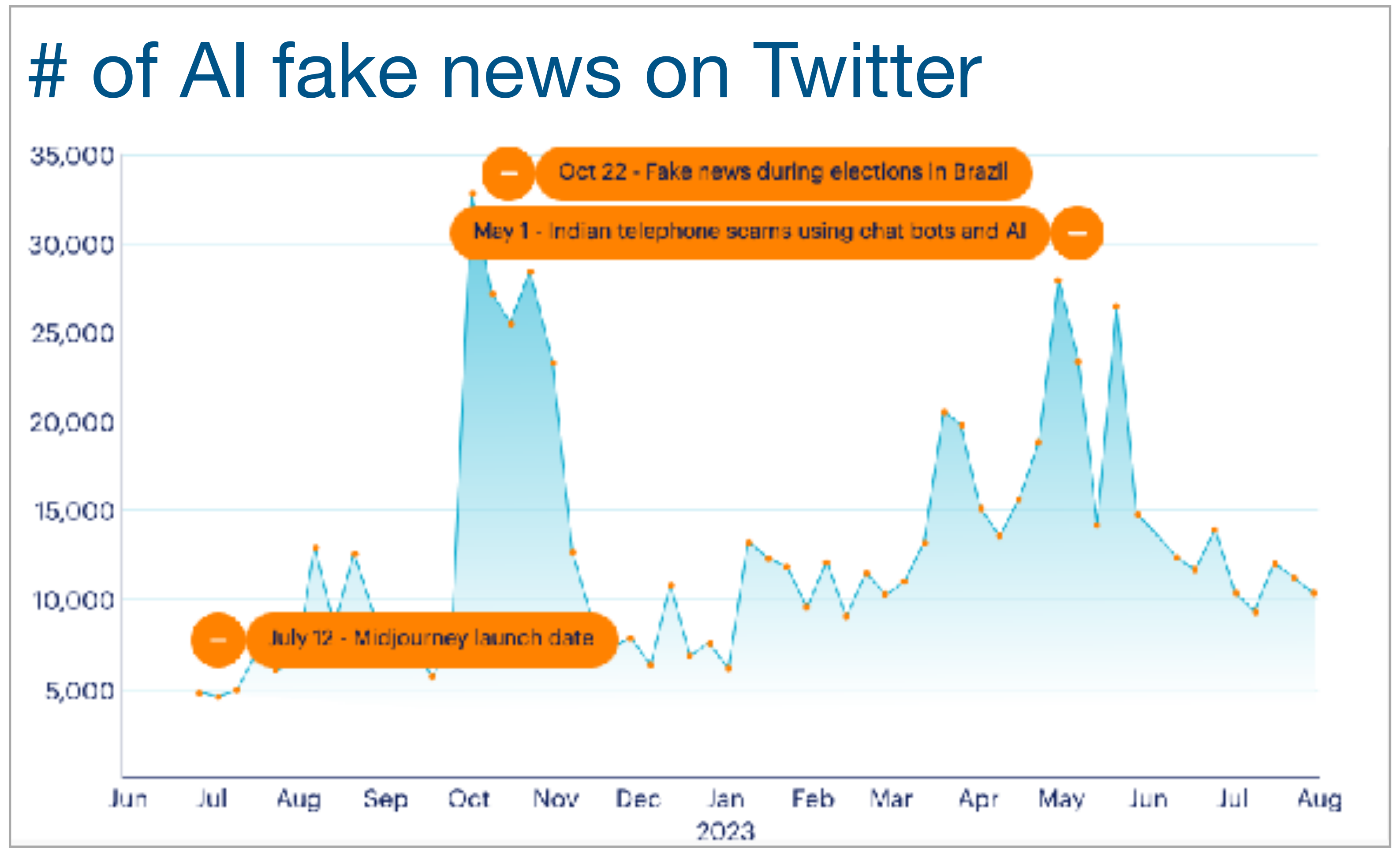
Challenges in AI Safety

Misuse of AI-generated content

[Source] Netskope.com and public sources.

Challenges in AI Safety

Misuse of AI-generated content



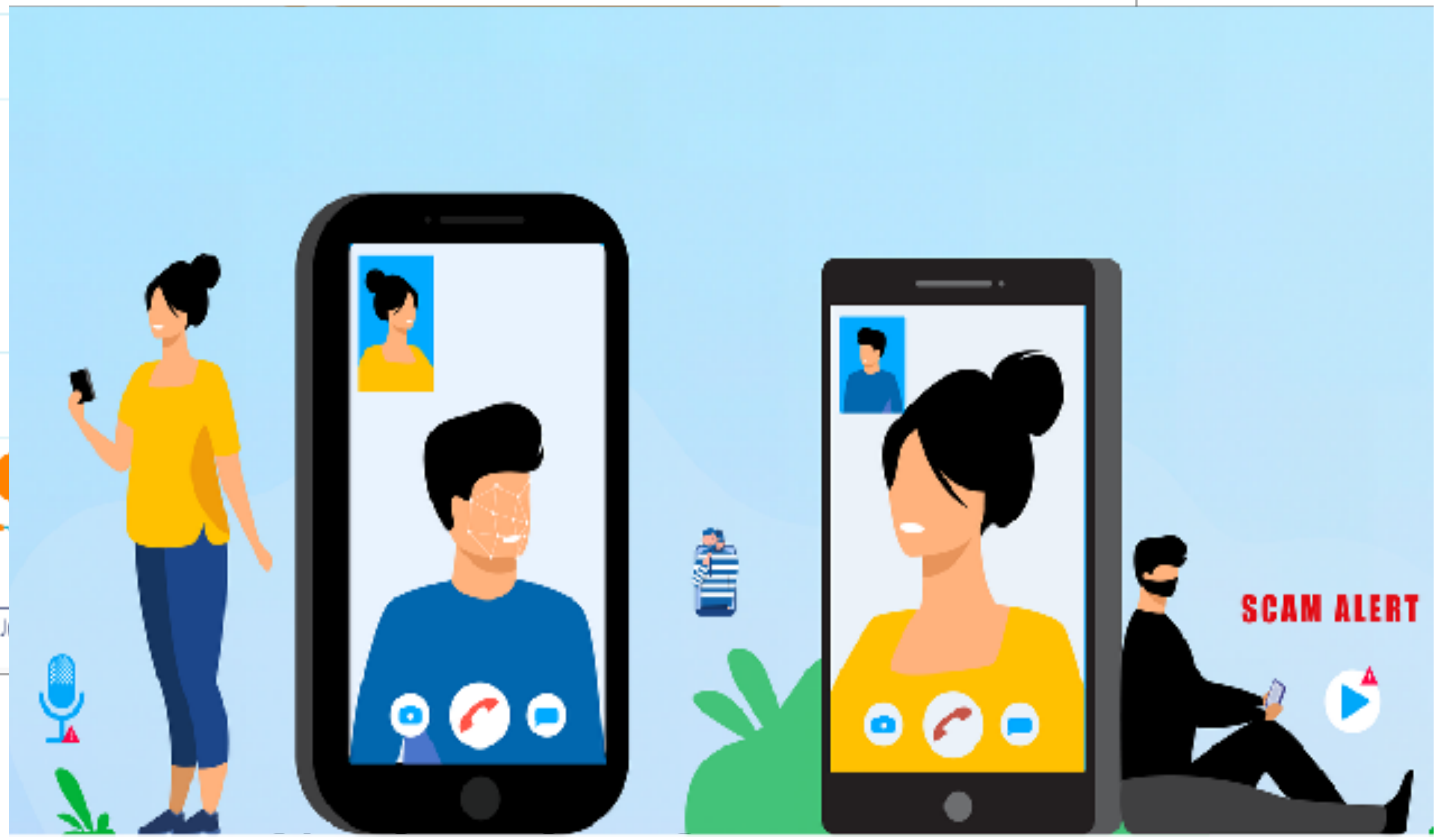
Fake news

[Source] Netskope.com and public sources.

Challenges in AI Safety

Misuse of AI-generated content

of AI fake news on Twitter



AI scams

[Source] Netskope.com and public sources.

Challenges in AI Safety

Misuse of AI-generated content

Plagiarism

[Source] Netskope.com and public sources.

Challenges in AI Safety

Misuse of AI-generated content

Data Pollution

Plagiarism

[Source] Netskope.com and public sources.

Challenges in AI Safety

Misuse of AI-generated content

Plagiarism

Data Pollution

Tons of AI-generate data over the internet

[Source] Netskope.com and public sources.

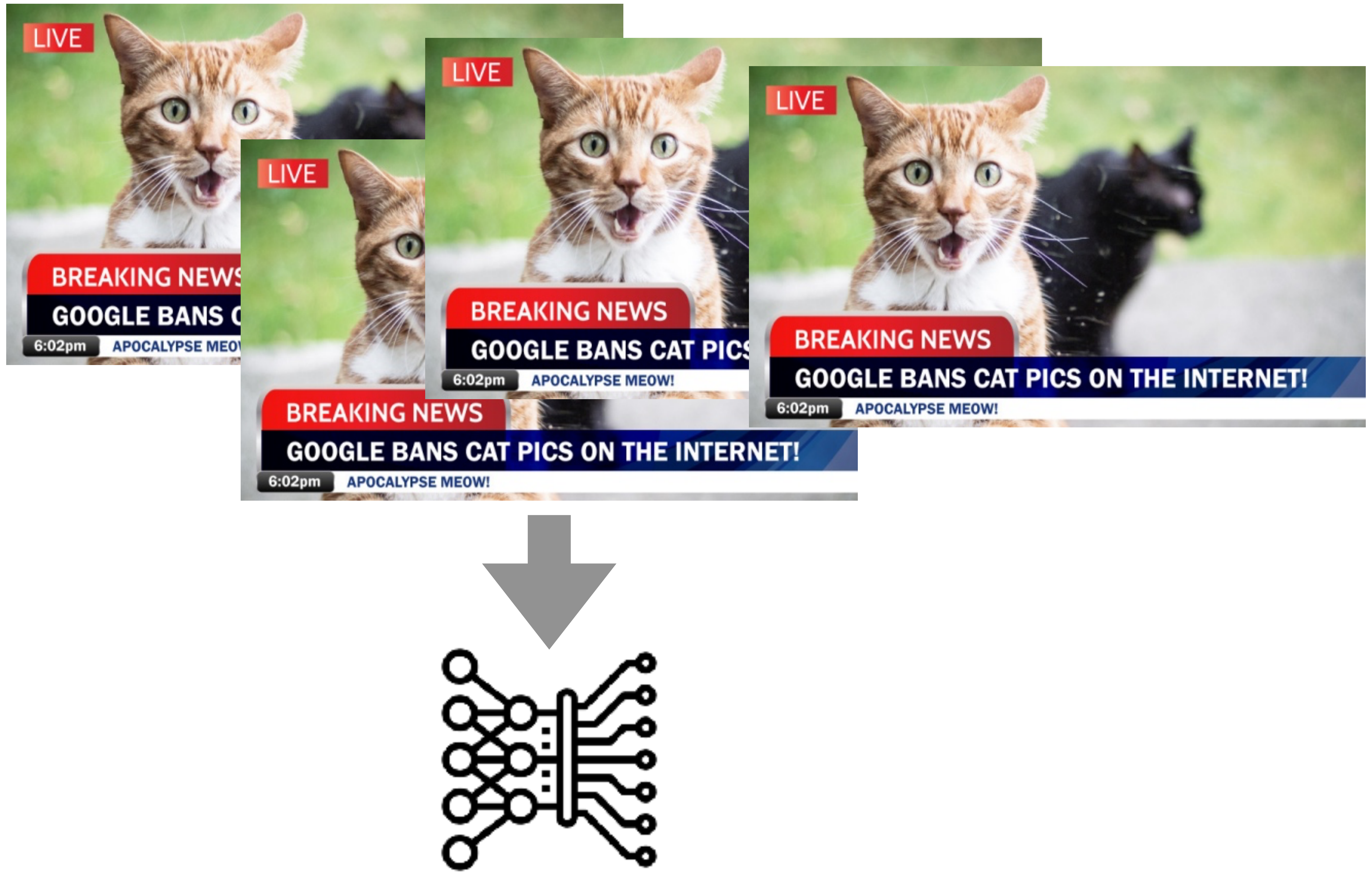
Challenges in AI Safety

Misuse of AI-generated content

Plagiarism

Data Pollution

Tons of AI-generate data over the internet



[Source] Netskope.com and public sources.

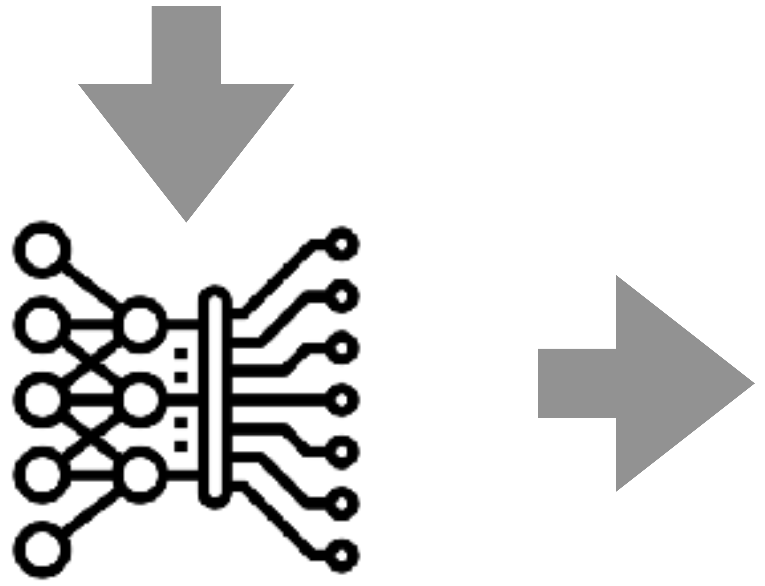
Challenges in AI Safety

Misuse of AI-generated content

Plagiarism

Data Pollution

Tons of AI-generate data over the internet



???
No cats anymore?

[Source] Netskope.com and public sources.

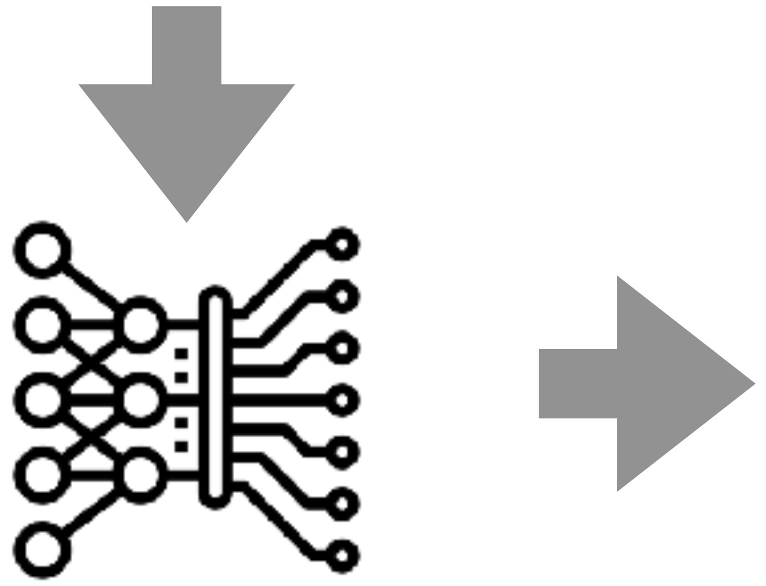
Challenges in AI Safety

Misuse of AI-generated content

Plagiarism

Data Pollution

Tons of AI-generate data over the internet



???
No cats anymore?

collapse...

[Source] Netskope.com and public sources.

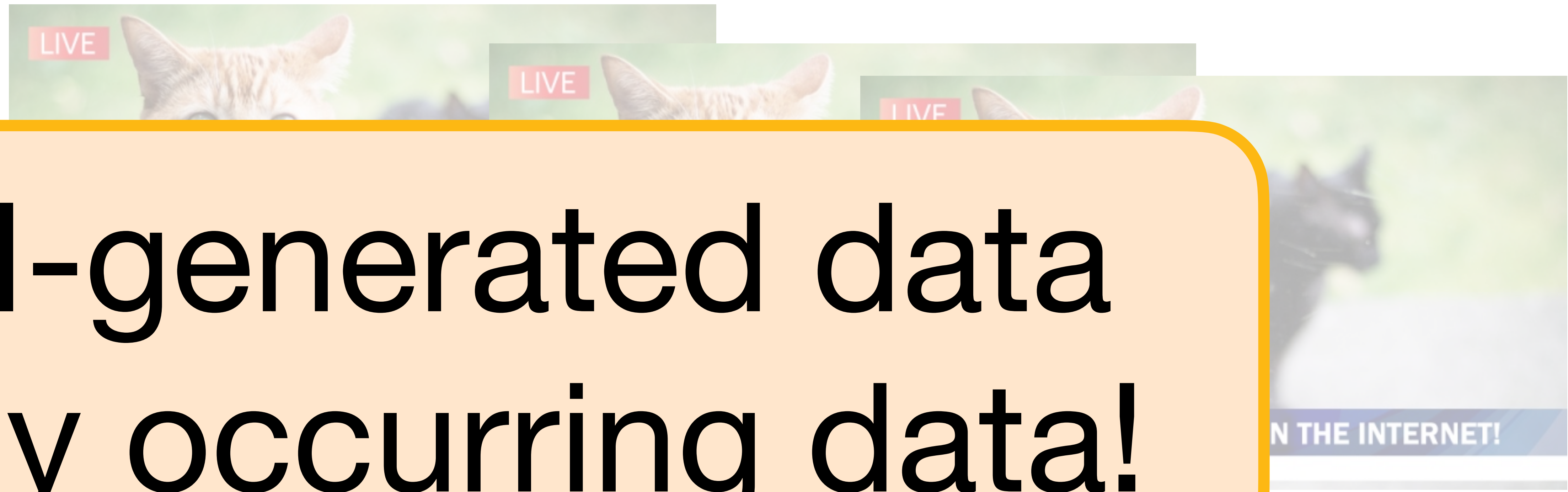
Challenges in AI Safety

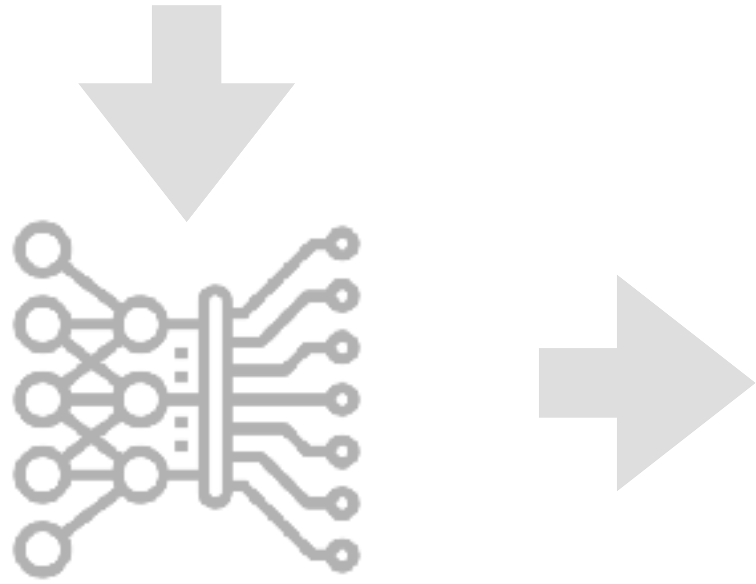
Misuse of AI-generated content

Plagiarism

Data Pollution

Tons of AI-generate data over the internet





???
No cats anymore?

collapse…

We must distinguish AI-generated data from authentic, naturally occurring data!

[Source] Netskope.com and public sources.

Identify AI-generated Text

Possible solutions?

Identify AI-generated Text

Possible solutions?

- By observation:

Identify AI-generated Text

Possible solutions?

“Here’s the revised version of your…”, *“Best regards,[Your Name]”* :-D

Identify AI-generated Text

Possible solutions?

- Metadata <—easy to remove

Metadata

File name: Dataset
Author: GPT
Location: Ithaca
Created: Jan 01, 2025

Identify AI-generated Text

Possible solutions?

- Giant database to store all AI-generated content <—storage? privacy?

Identify AI-generated Text

Possible solutions?

- Discriminator models: GPTZero DetectGPT Copyleaks pangramlabs …

Identify AI-generated Text

Possible solutions?

<—high prob of falsely alarming human-written text

Identify AI-generated Text

Possible solutions?

- **Watermarking: inserting a signal into LLM predicted tokens**

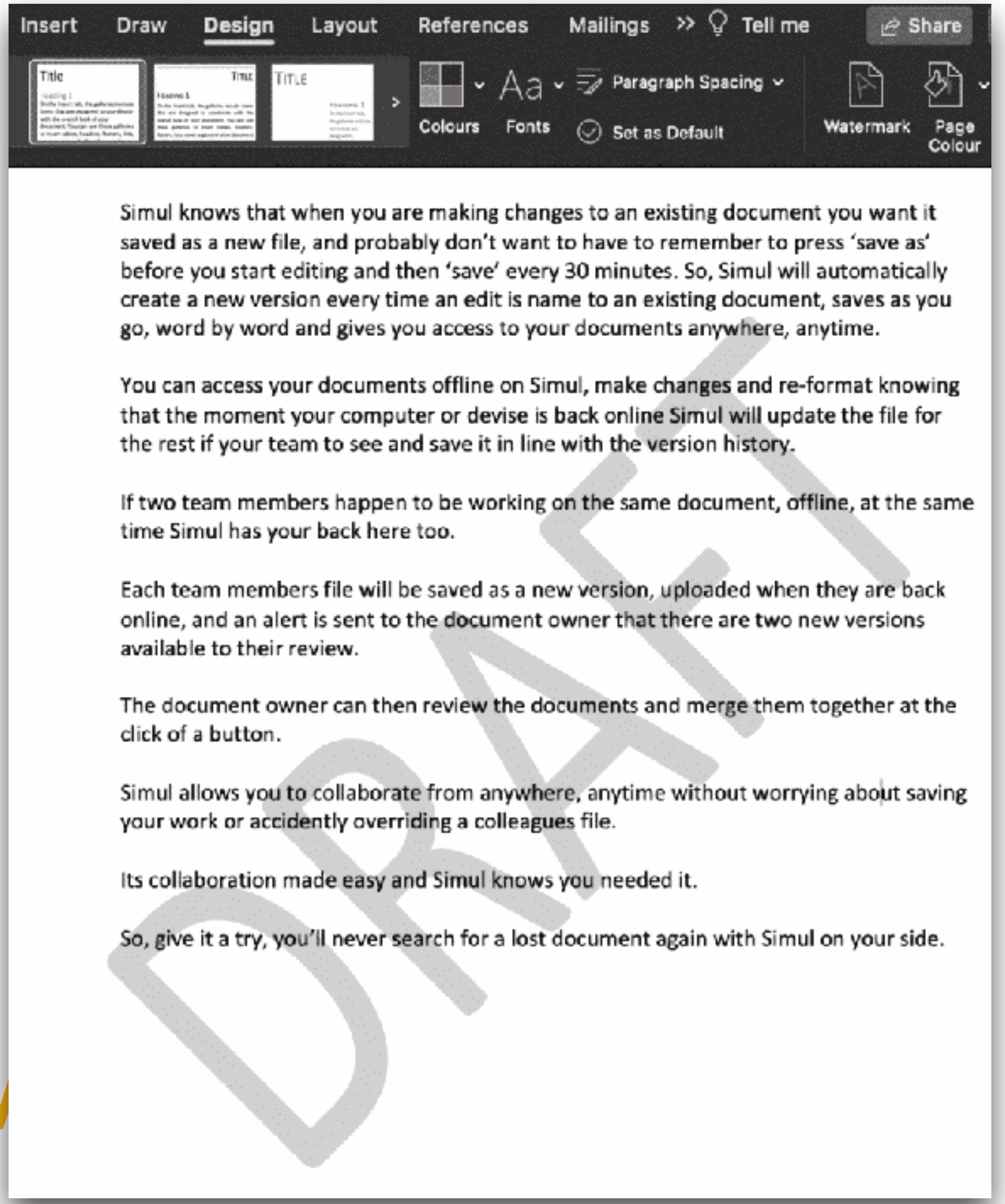
Identify AI-generated Text

Possible solutions?

- **Watermarking: inserting a signal into LLM predicted tokens**

Identify AI-generated Text

Possible solutions?



Simul knows that when you are making changes to an existing document you want it saved as a new file, and probably don't want to have to remember to press 'save as' before you start editing and then 'save' every 30 minutes. So, Simul will automatically create a new version every time an edit is name to an existing document, saves as you go, word by word and gives you access to your documents anywhere, anytime.

You can access your documents offline on Simul, make changes and re-format knowing that the moment your computer or devise is back online Simul will update the file for the rest if your team to see and save it in line with the version history.

If two team members happen to be working on the same document, offline, at the same time Simul has your back here too.

Each team members file will be saved as a new version, uploaded when they are back online, and an alert is sent to the document owner that there are two new versions available to their review.

The document owner can then review the documents and merge them together at the click of a button.

Simul allows you to collaborate from anywhere, anytime without worrying about saving your work or accidently overriding a colleagues file.

Its collaboration made easy and Simul knows you needed it.

So, give it a try, you'll never search for a lost document again with Simul on your side.

- **Watermarking: inserting a signal into LLM**

Identify AI-generated Text

Possible solutions?

- **Watermarking: inserting a signal into LLM predicted tokens**

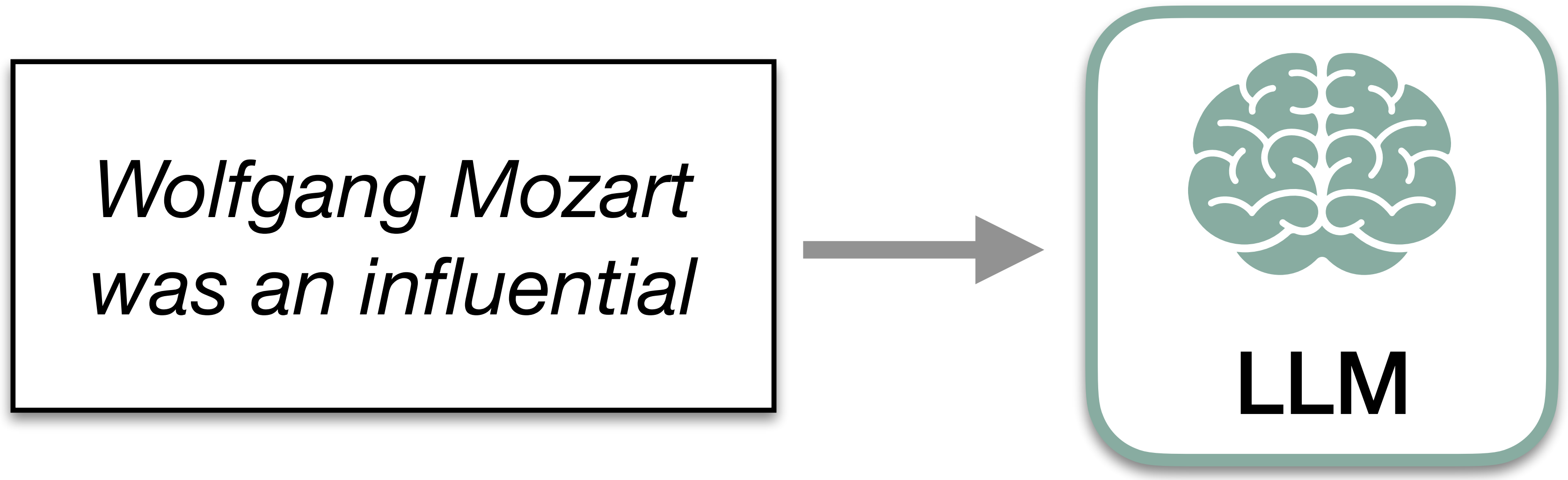
Identify AI-generated Text

Possible solutions?

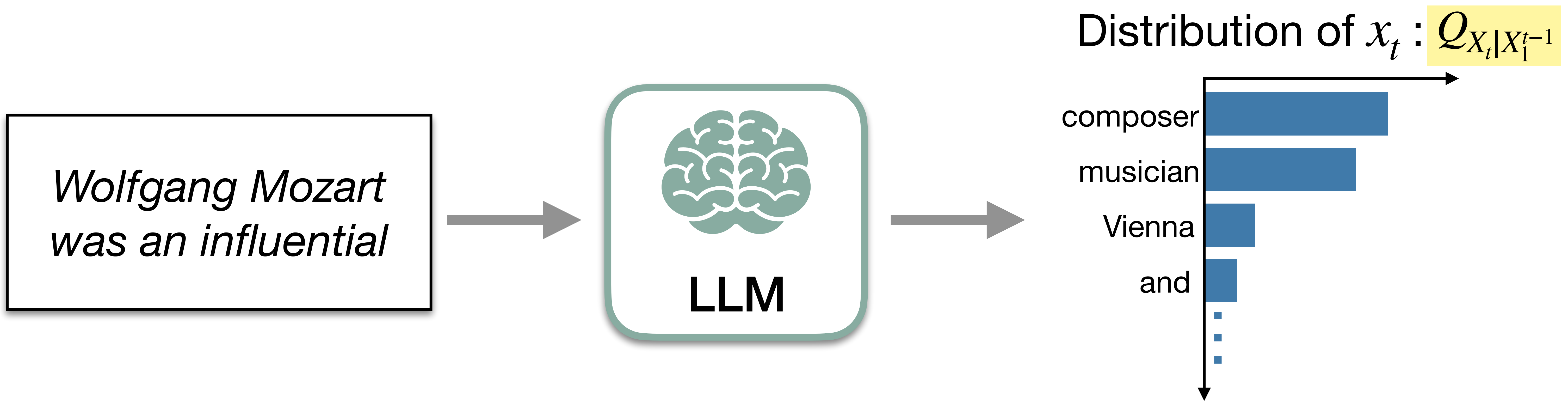
- **Watermarking: inserting a signal into LLM predicted tokens**

A Framework for LLM Watermark Generation

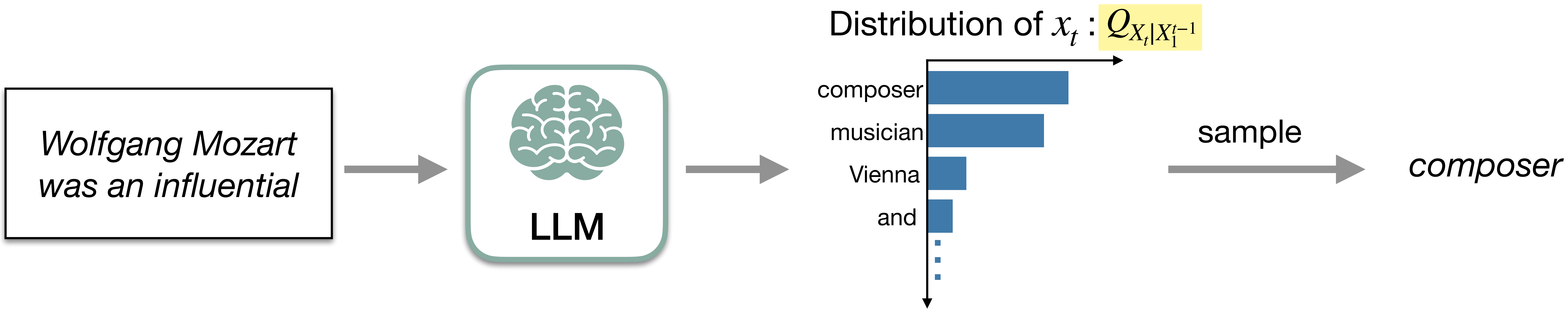
A Framework for LLM Watermark Generation



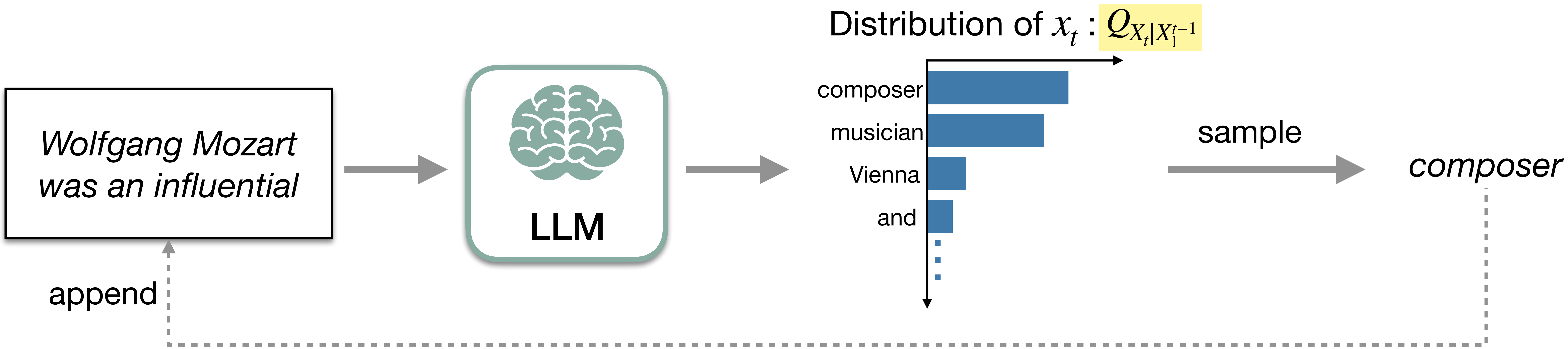
A Framework for LLM Watermark Generation



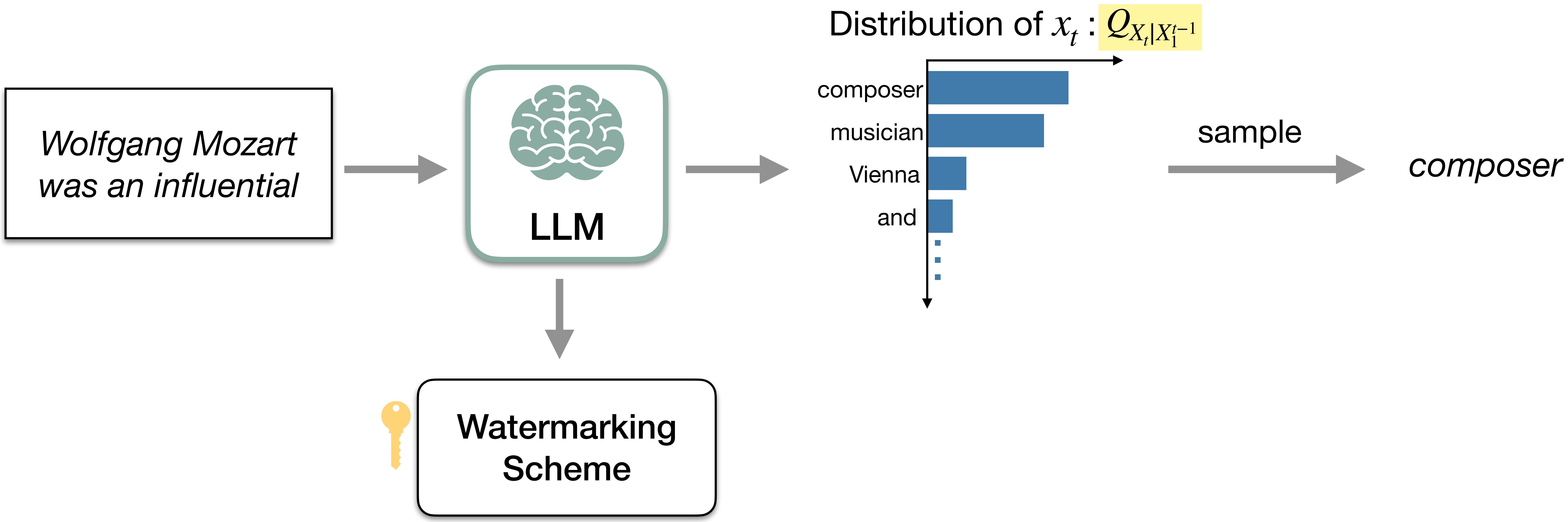
A Framework for LLM Watermark Generation



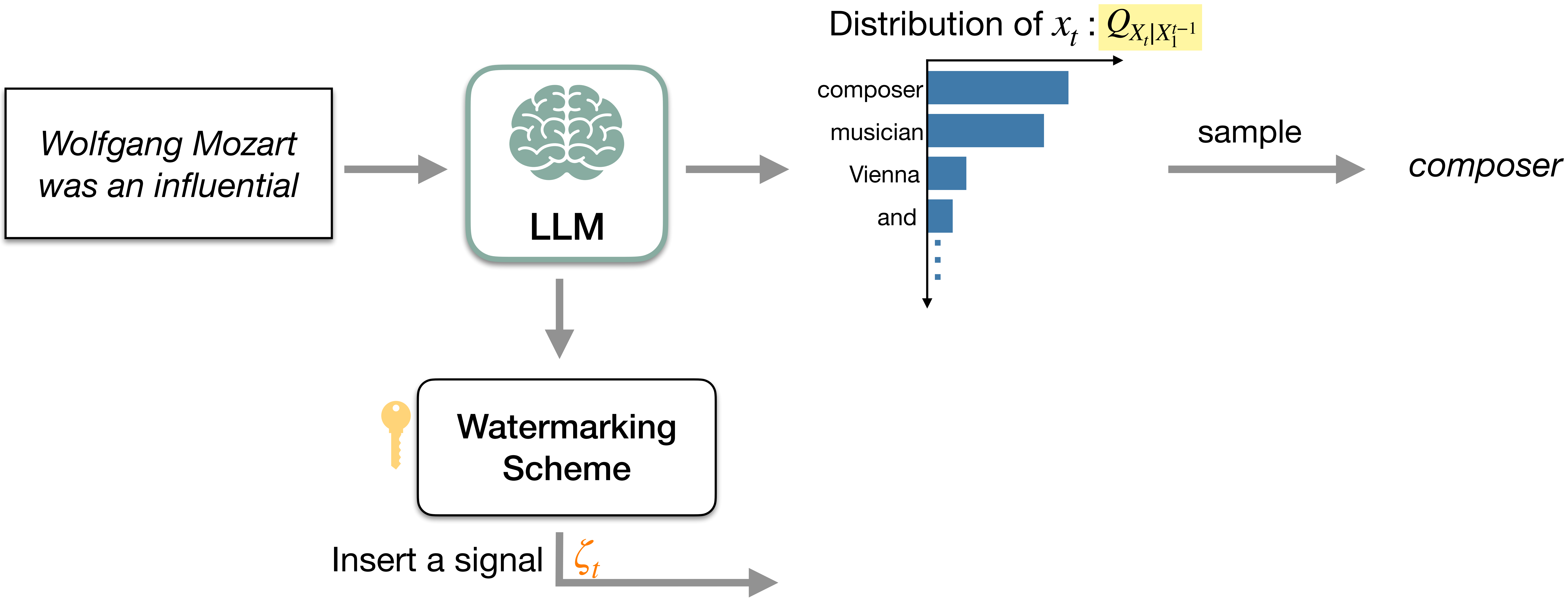
A Framework for LLM Watermark Generation



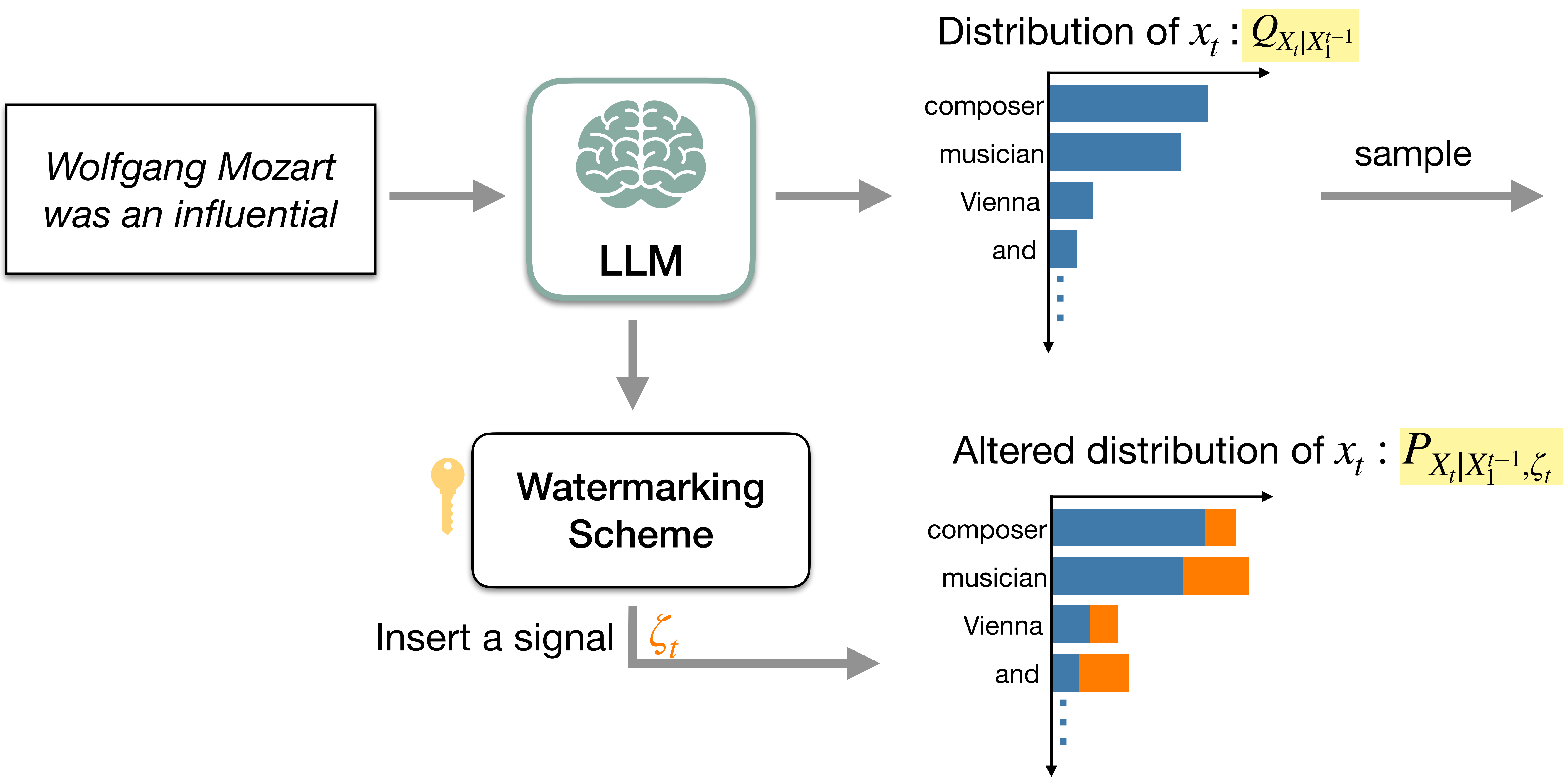
A Framework for LLM Watermark Generation



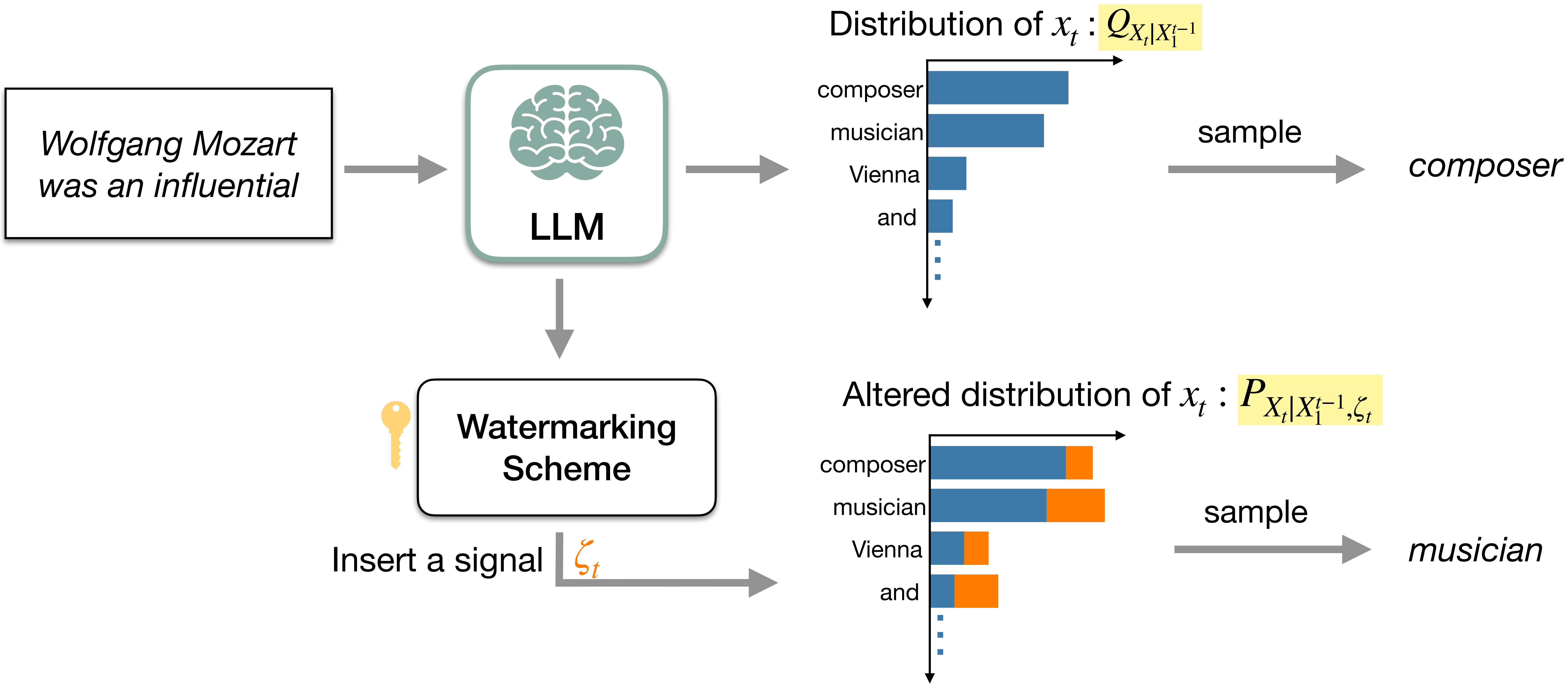
A Framework for LLM Watermark Generation



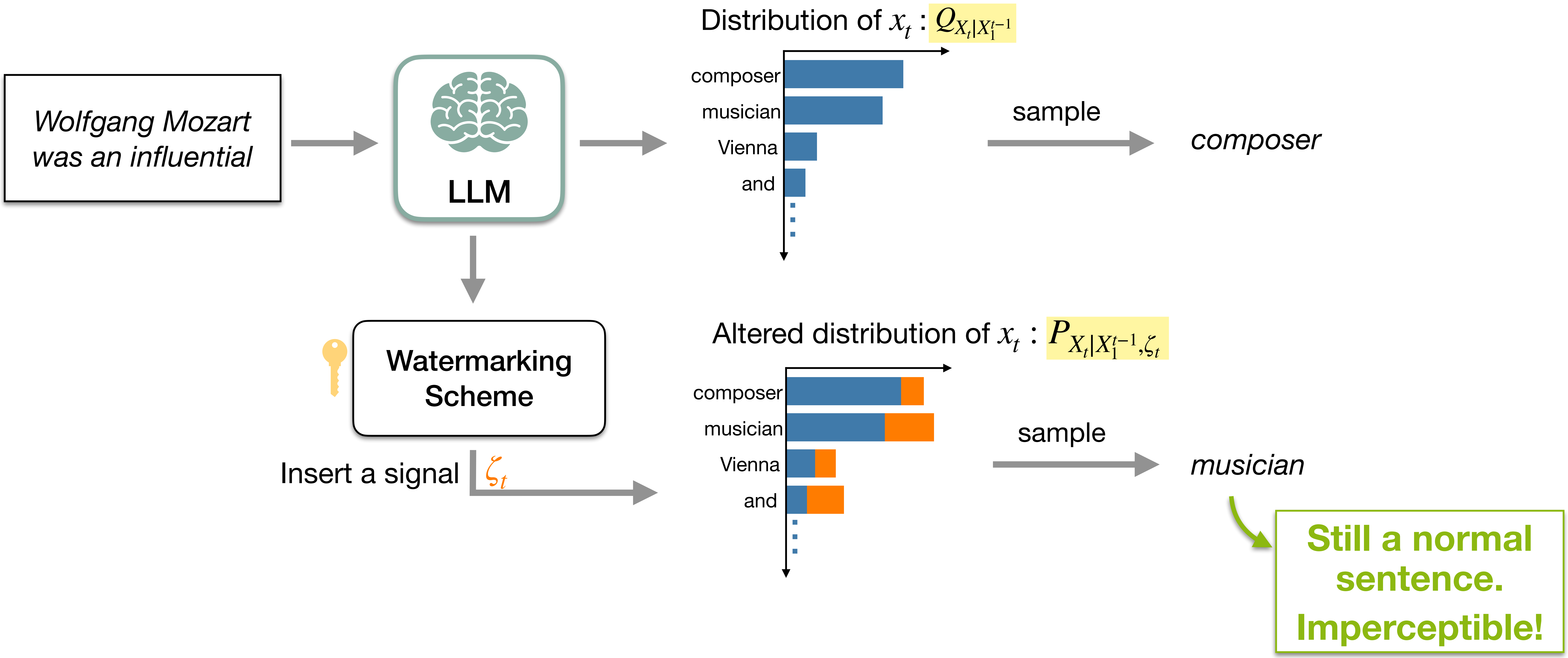
A Framework for LLM Watermark Generation



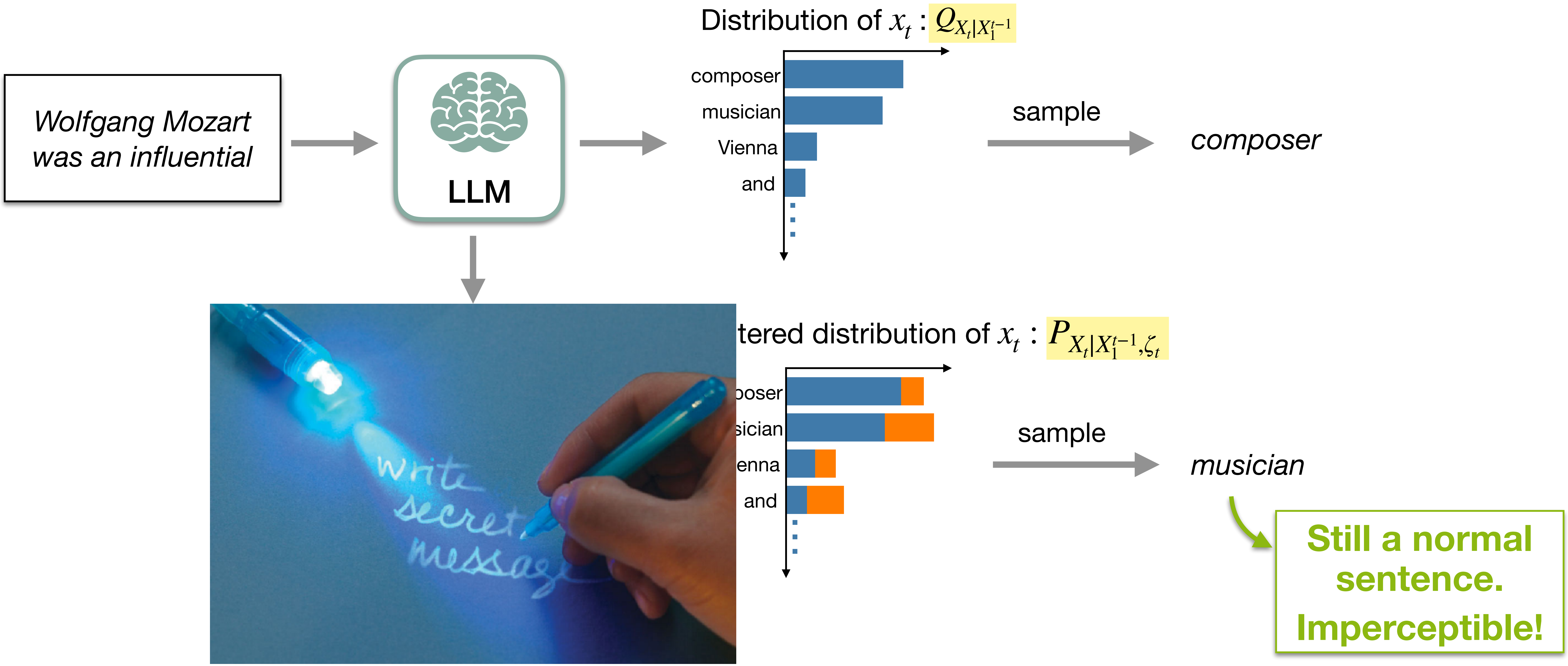
A Framework for LLM Watermark Generation



A Framework for LLM Watermark Generation

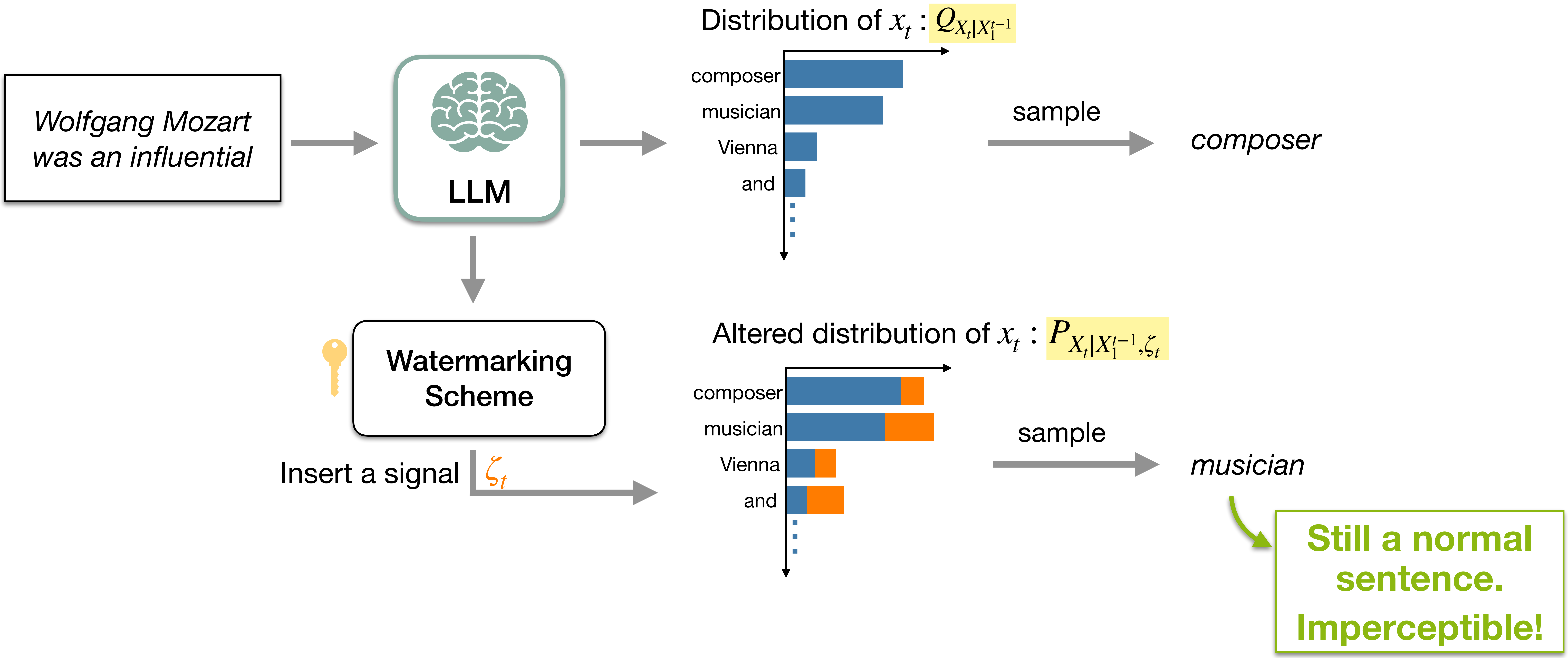


A Framework for LLM Watermark Generation

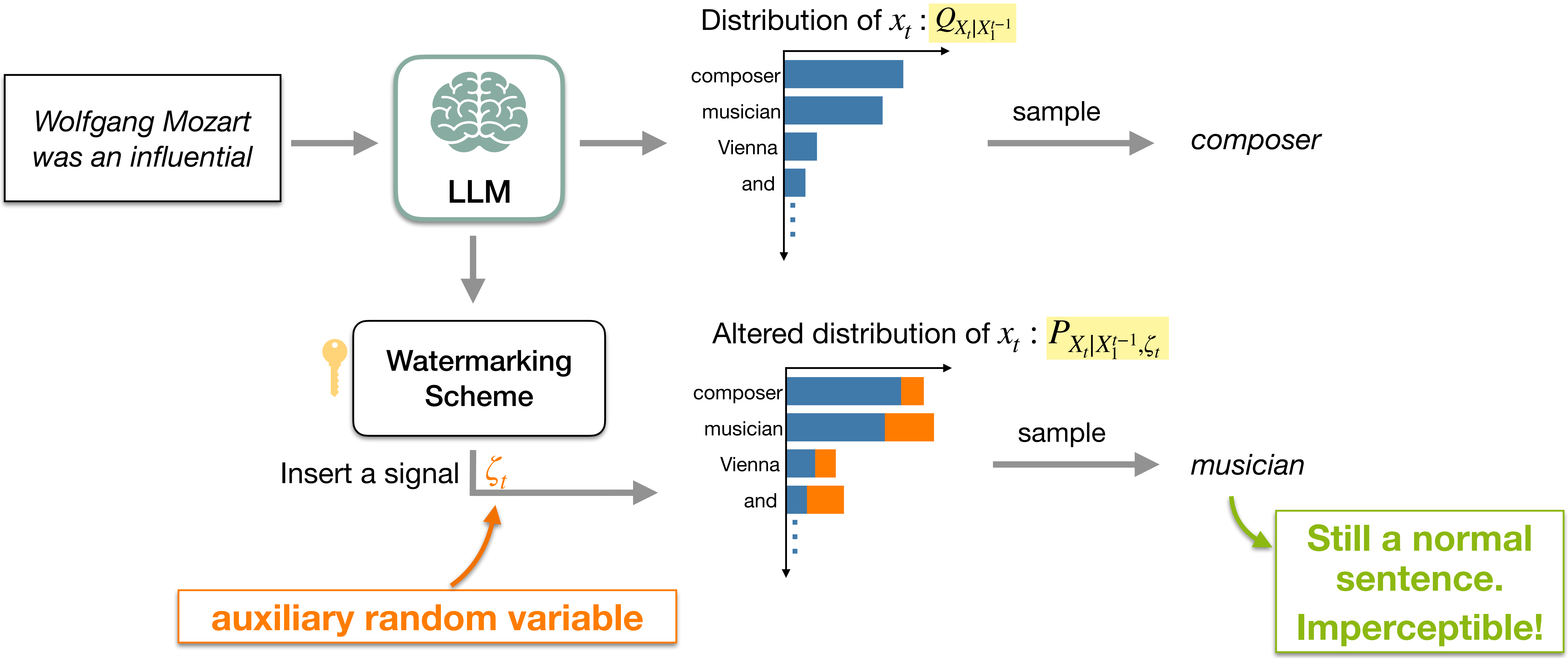


Like invisible Ink (Steganography)

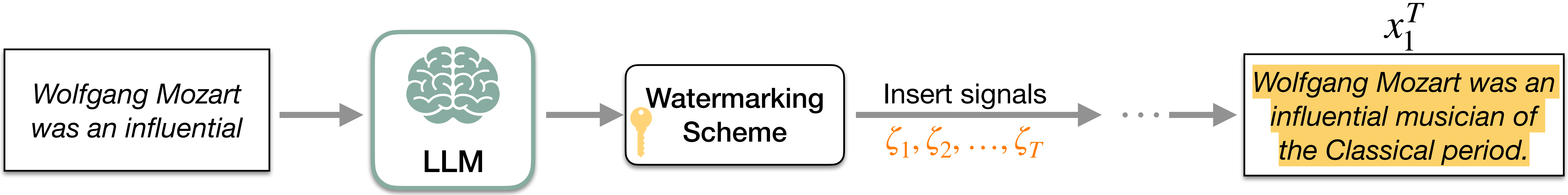
A Framework for LLM Watermark Generation



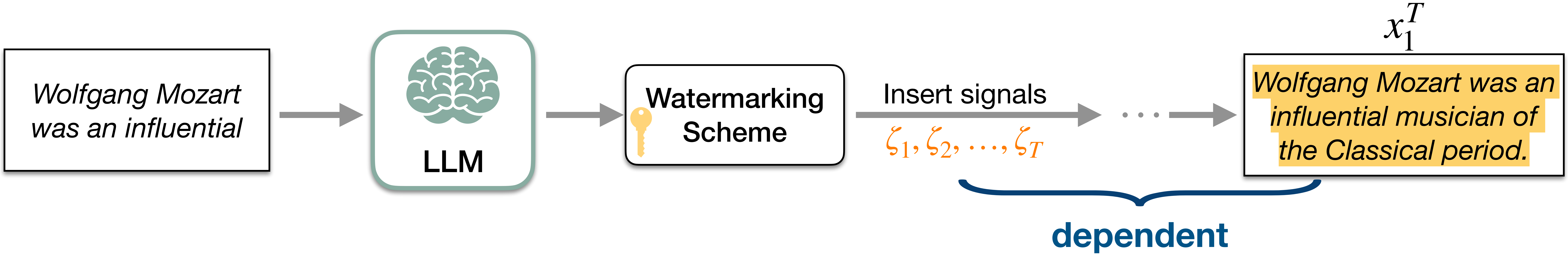
A Framework for LLM Watermark Generation



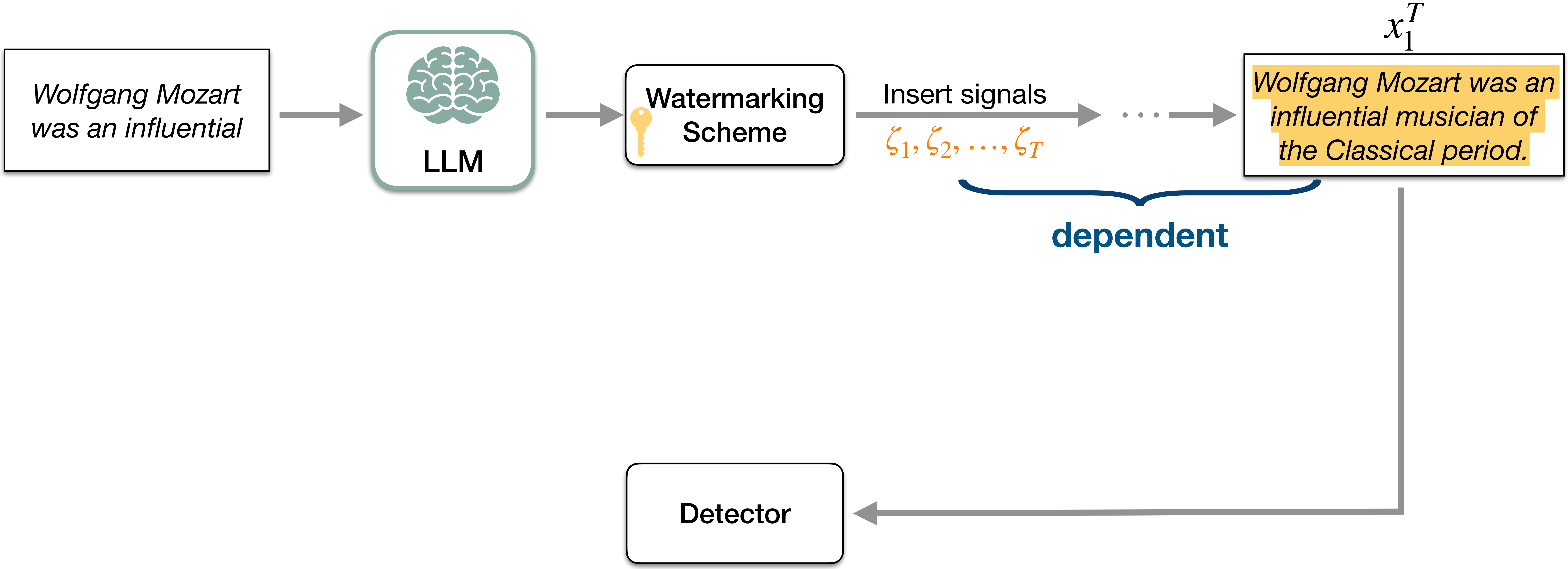
Hypothesis Testing for LLM Watermark Detection



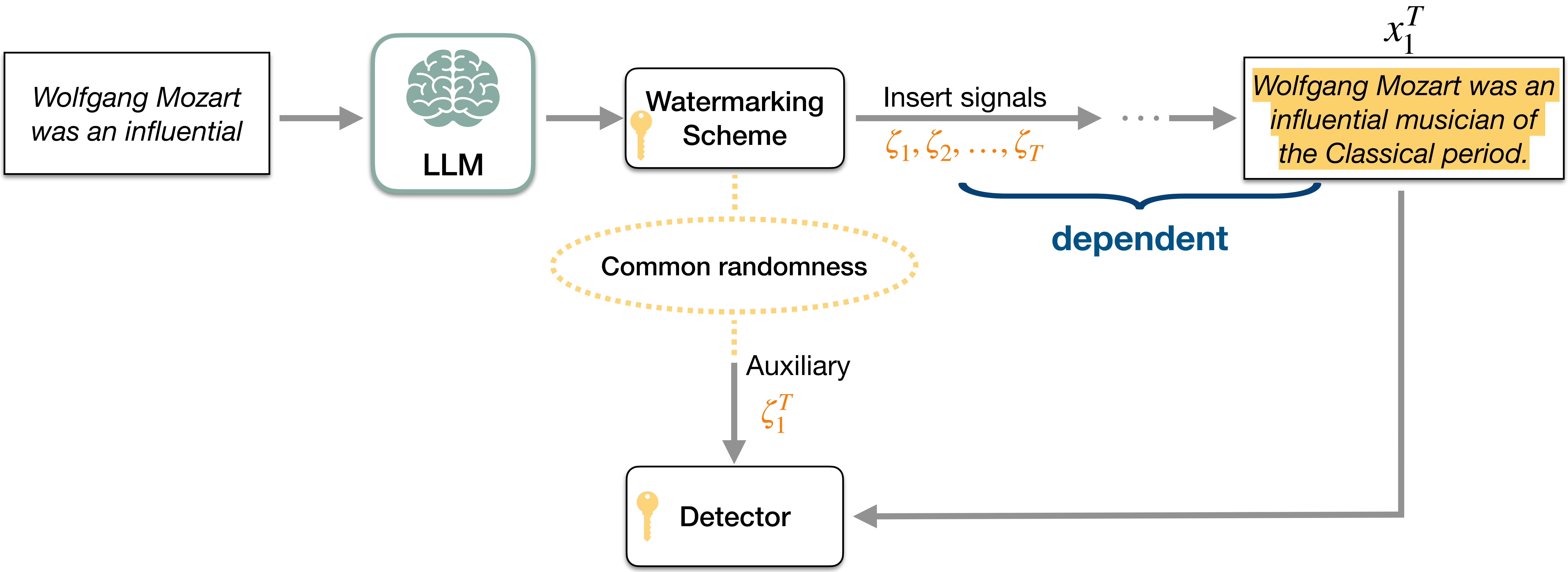
Hypothesis Testing for LLM Watermark Detection



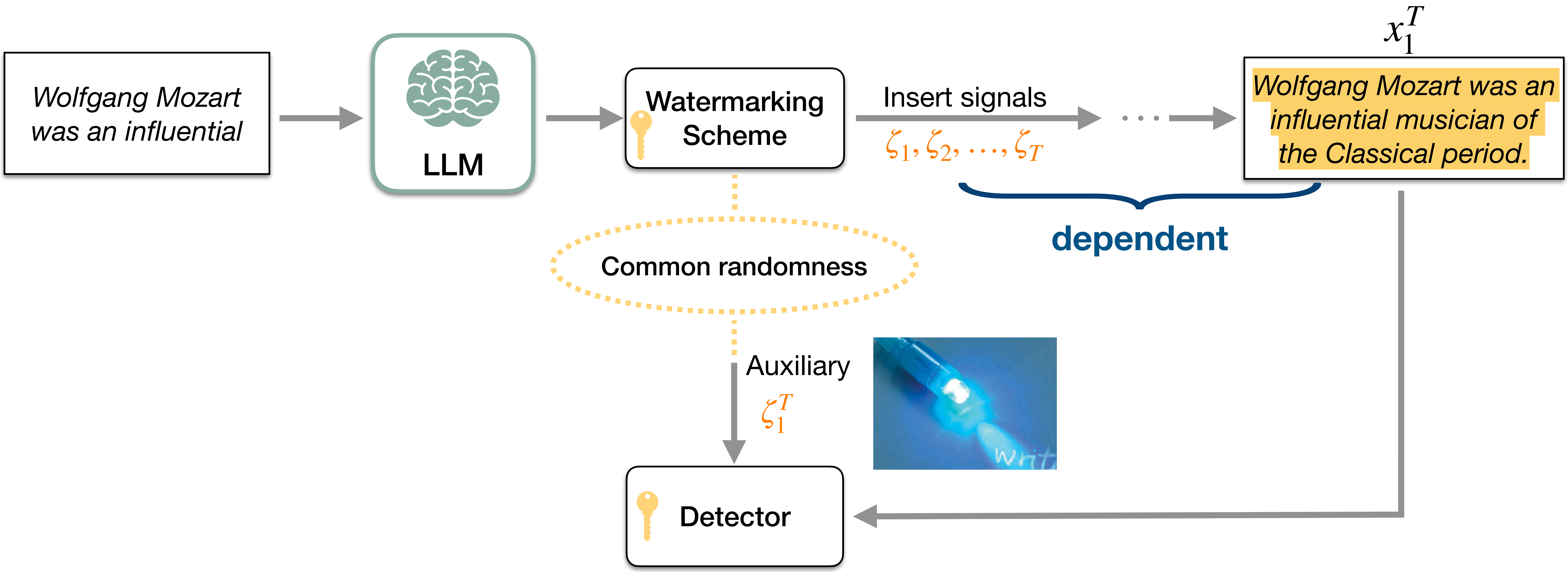
Hypothesis Testing for LLM Watermark Detection



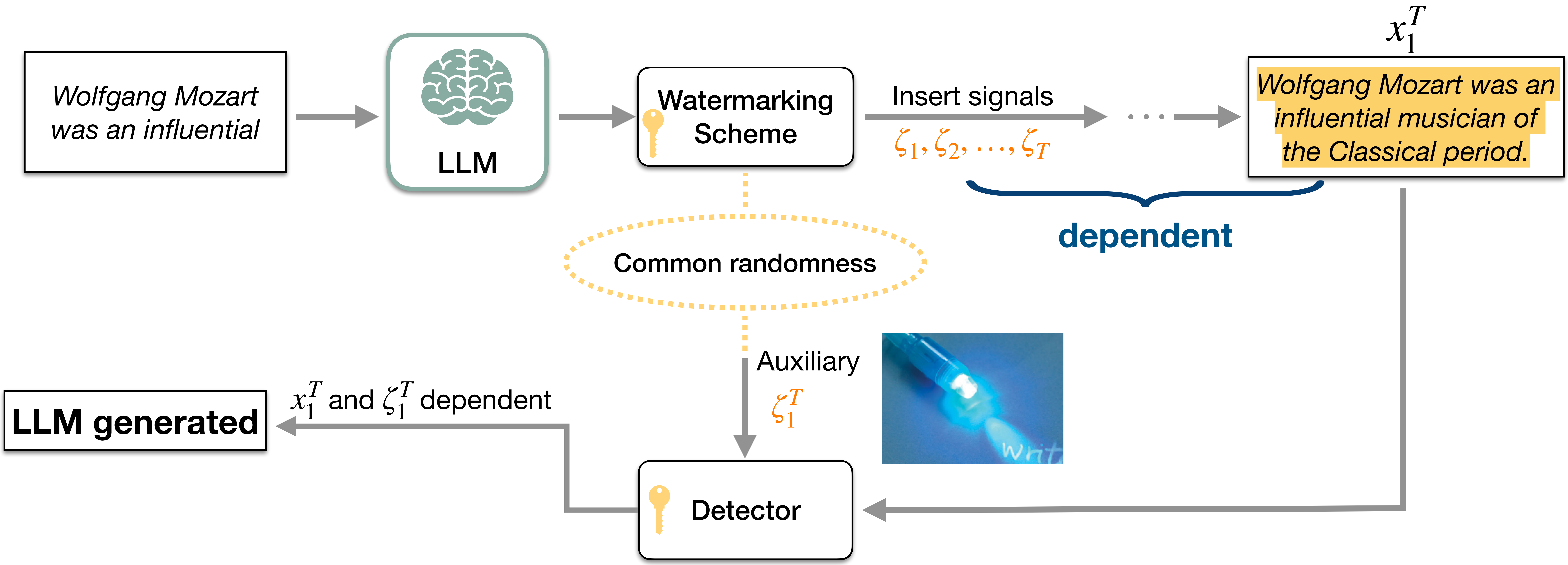
Hypothesis Testing for LLM Watermark Detection



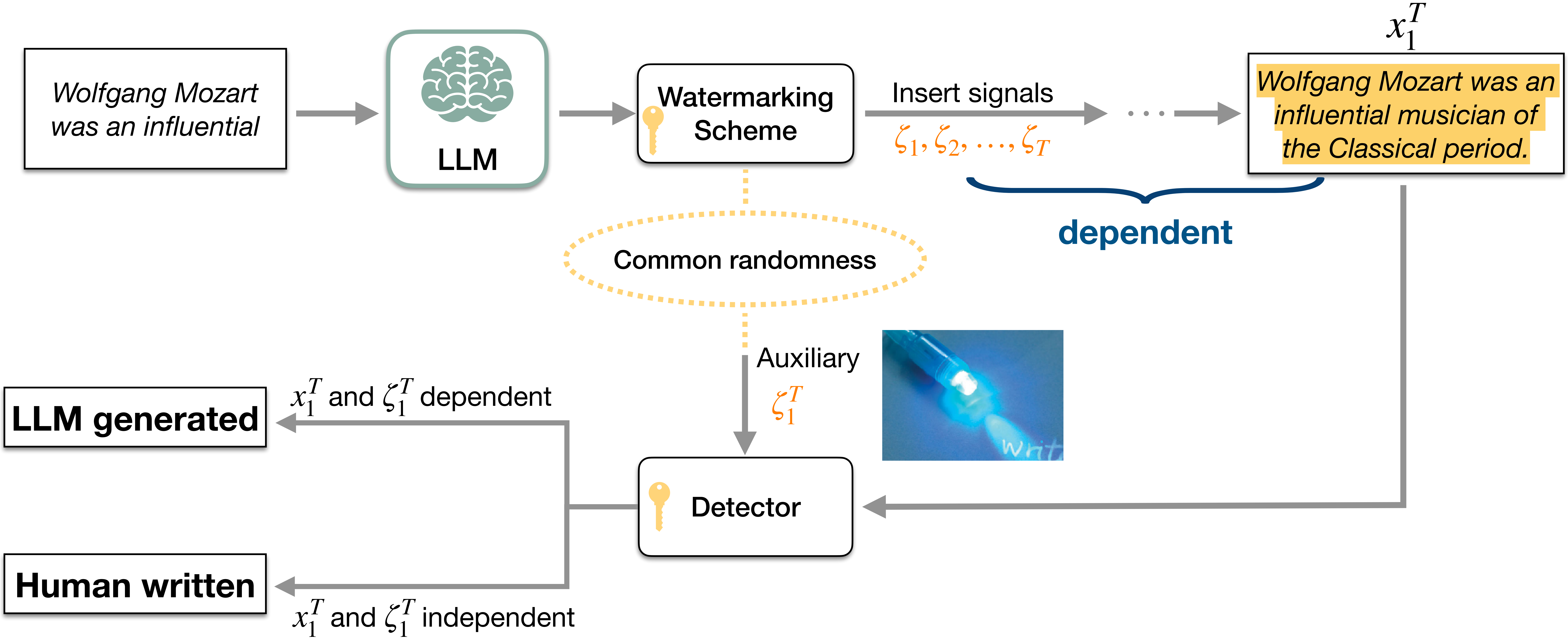
Hypothesis Testing for LLM Watermark Detection



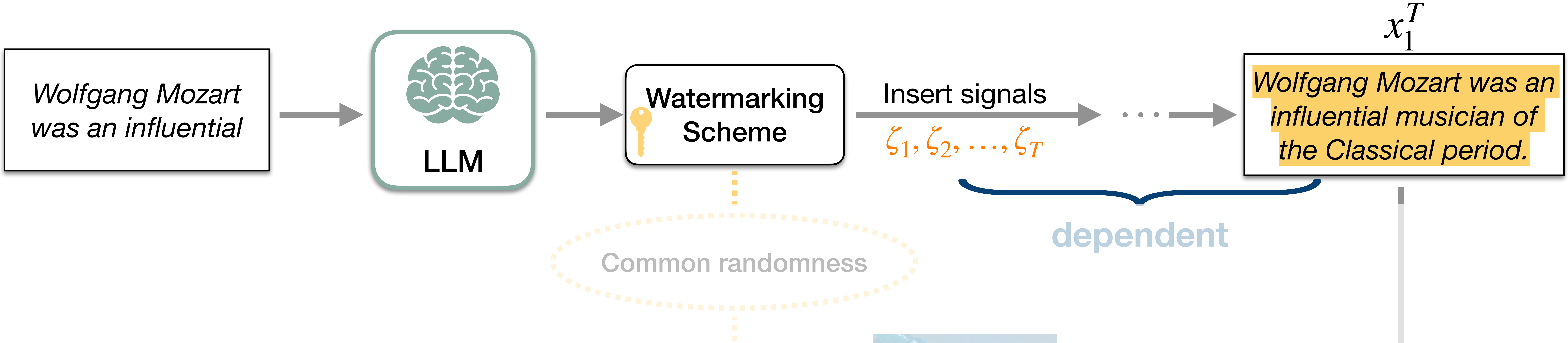
Hypothesis Testing for LLM Watermark Detection



Hypothesis Testing for LLM Watermark Detection



Hypothesis Testing for LLM Watermark Detection



Watermark Detection $\Longrightarrow$ Hypothesis Testing:

$$\mathrm{H}_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$\mathrm{H}_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

LLM

Hum

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$\mathrm{H}_0 : X_1^T$ is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

$\mathrm{H}_1 : X_1^T$ is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Performance metric:

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$H_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$H_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Watermarking scheme

Performance metric:

Detector γ $\begin{cases} H_0 : \text{Human} \\ H_1 : \text{LLM} \end{cases}$

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$H_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$H_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Watermarking scheme

Performance metric:

Reality

Detector γ $\begin{cases} H_0 : \text{Human} \\ H_1 : \text{LLM} \end{cases}$

	H_0 : Human	H_1 : LLM
H_0 : Human		
H_1 : LLM		

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Performance metric:

Reality

Detector γ	H_0 : Human	H_1 : LLM
H_0 : Human	✅	
H_1 : LLM	Type-I error (false alarm) $\beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T})$	

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Performance metric:

Reality

Detector γ $\begin{cases} \mathrm{H}_0 : \text{Human} \\ \mathrm{H}_1 : \text{LLM} \end{cases}$

	H_0 : Human	H_1 : LLM
H_0 : Human	✅	Type-II error (miss detection) $\beta_1(\gamma, P_{X_1^T, \zeta_1^T})$
H_1 : LLM	Type-I error (false alarm) $\beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T})$	✅

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$\mathrm{H}_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Watermarking scheme

Performance metric:

Detector γ $\begin{cases} \mathrm{H}_0 : \text{Human} \\ \mathrm{H}_1 : \text{LLM} \end{cases}$

Reality

	H_0 : Human	H_1 : LLM
H_0 : Human	✅	Type-II error (miss detection) $\min\ \beta_1(\gamma, P_{X_1^T, \zeta_1^T})$
H_1 : LLM	Type-I error (false alarm) $\beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$	✅

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$\mathrm{H}_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Watermarking scheme

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

scheme

Other criteria for LLM watermarking?

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$H_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$H_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

scheme

Other criteria for LLM watermarking?

$\Longrightarrow$ **Text Quality!**

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$\mathrm{H}_0 : X_1^T$ is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

$\mathrm{H}_1 : X_1^T$ is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

scheme

Other criteria for LLM watermarking?

$\Longrightarrow$ **Text Quality!**

$\Longrightarrow$ **Indistinguishable from unwatermarked**

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T **is human written, i.e.,** $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T **is LLM generated, i.e.,** $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

watermarked text distribution

$P_{X_1^T}$

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Human/unwatermarked LLM

Watermarking scheme

watermarked text distribution $P_{X_1^T}$ **vs** original text distribution $Q_{X_1^T}$

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$\mathrm{H}_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Watermarking scheme

watermarked text distribution $P_{X_1^T}$ **vs** original text distribution $Q_{X_1^T}$

Good text quality

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$\mathrm{H}_0 : X_1^T$ **is human written, i.e.,** $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

$\mathrm{H}_1 : X_1^T$ **is LLM generated, i.e.,** $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

watermarked text distribution $P_{X_1^T}$ **vs** original text distribution $Q_{X_1^T}$

Good text quality $\Longleftrightarrow$ $\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

watermarked text distribution $P_{X_1^T}$ **vs** original text distribution $Q_{X_1^T}$

Good text quality $\Longleftrightarrow$ $\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$

(Distortion Level)

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

watermarked text distribution $P_{X_1^T}$ vs original text distribution $Q_{X_1^T}$

Good text quality $\Longleftrightarrow$ $D(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$ (D can be any distortion metric)

ϵ: (Distortion Level)

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$\mathrm{H}_0 : X_1^T$ **is human written, i.e.,** $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

$\mathrm{H}_1 : X_1^T$ **is LLM generated, i.e.,** $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$\mathrm{H}_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Watermarking scheme

Trade-off:

Type-II error— —False alarm rate— —Distortion Level

$$\beta_1 - -\alpha - -\epsilon$$

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$\mathrm{H}_0 : X_1^T$ is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

$\mathrm{H}_1 : X_1^T$ is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$\mathrm{H}_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Watermarking scheme

Existing watermarking methods: **heuristic**

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$\mathrm{H}_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Watermarking scheme

Existing watermarking methods: **heuristic**

Example
[KGW-1, 2023]

No watermark

```
Extremely efficient on average term
lengths and word frequencies on
synthetic, microamount text (as little
as 25 words)
Very small and low-resource key/hash
(e.g., 140 bits per key is sufficient
for 99.999999999% of the Synthetic
Internet
```

With watermark

```
- minimal marginal probability for a
detection attempt.
- Good speech frequency and energy
rate reduction.
- messages indiscernible to humans.
- easy for humans to verify.
```

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

$$\mathrm{H}_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

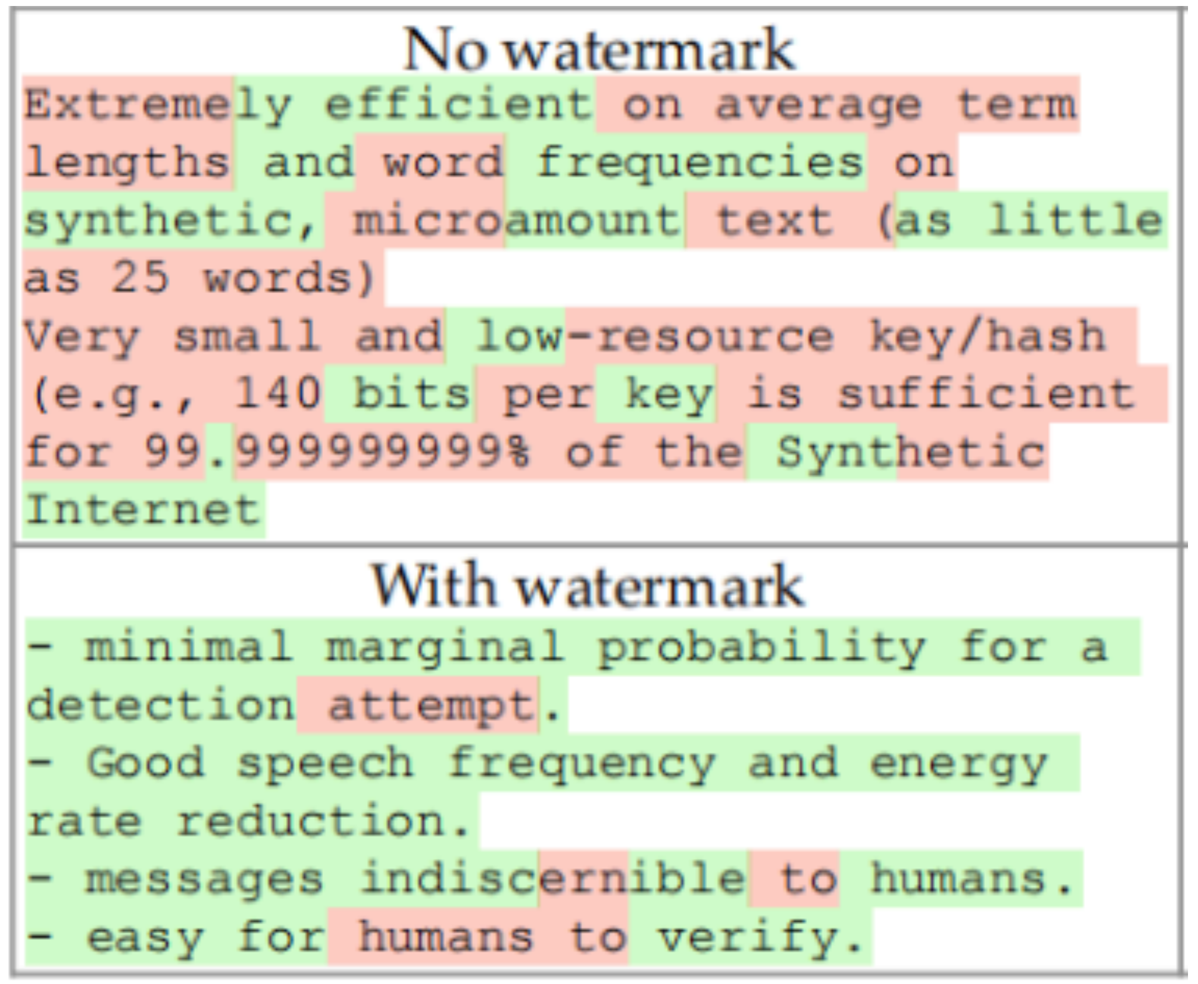
$$\mathrm{H}_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Human/unwatermarked LLM ($Q_{X_1^T}$)

Watermarking scheme ($P_{\zeta_1^T}$, $P_{X_1^T, \zeta_1^T}$)

Existing watermarking methods: **heuristic**

Example
[KGW-1, 2023]


No watermark
Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words)
Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet

With watermark
- minimal marginal probability for a detection attempt.
- Good speech frequency and energy rate reduction.
- messages indiscernible to humans.
- easy for humans to verify.

a word ← randomly assign green/red color

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

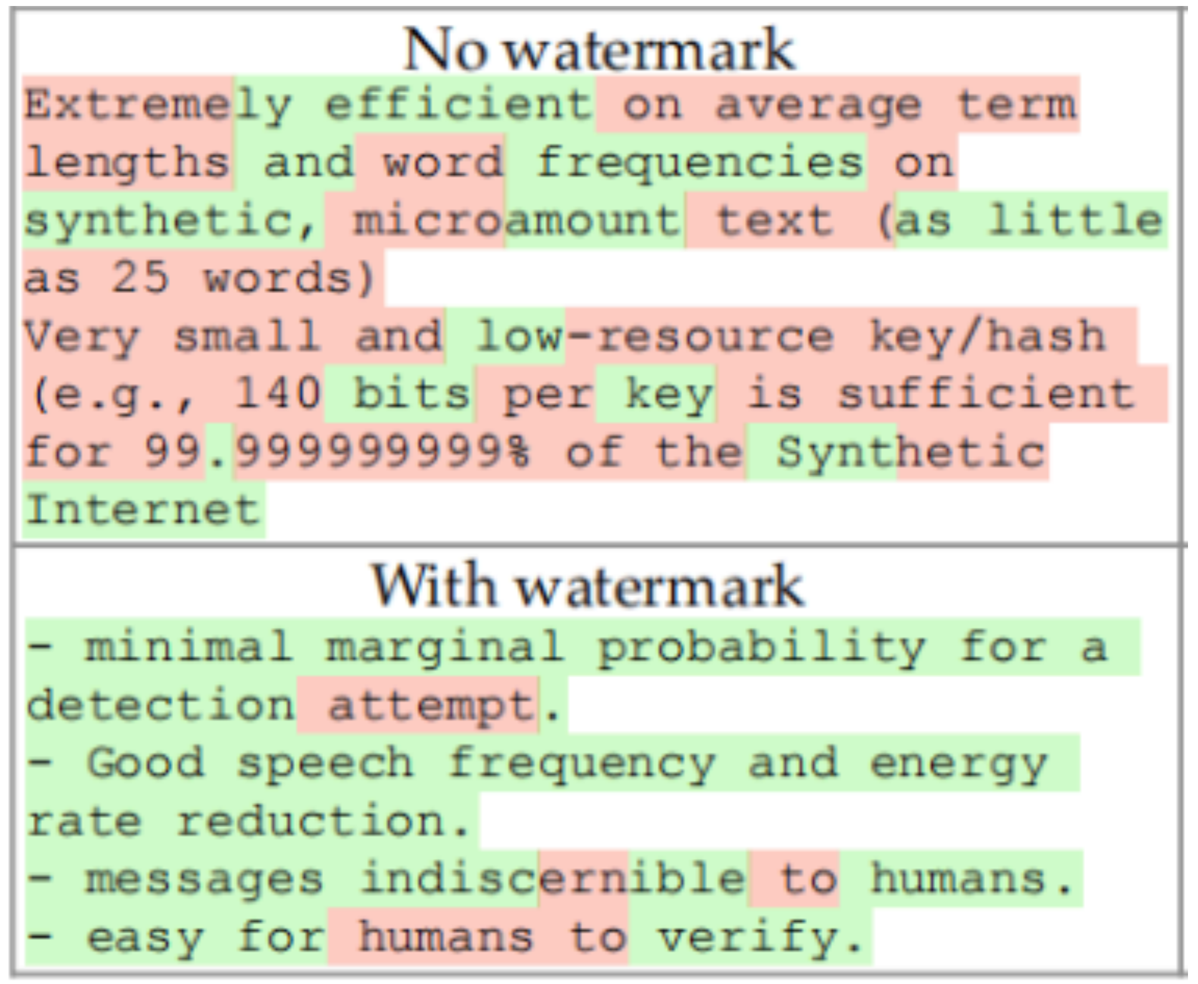
$$\mathrm{H}_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$\mathrm{H}_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Watermarking scheme

Existing watermarking methods: **heuristic**

Example
[KGW-1, 2023]


No watermark

Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words)
Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet

With watermark

- minimal marginal probability for a detection attempt.
- Good speech frequency and energy rate reduction.
- messages indiscernible to humans.
- easy for humans to verify.

a word←randomly assign green/red color

Green word: <u>increase</u> sampling probability

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

$$H_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

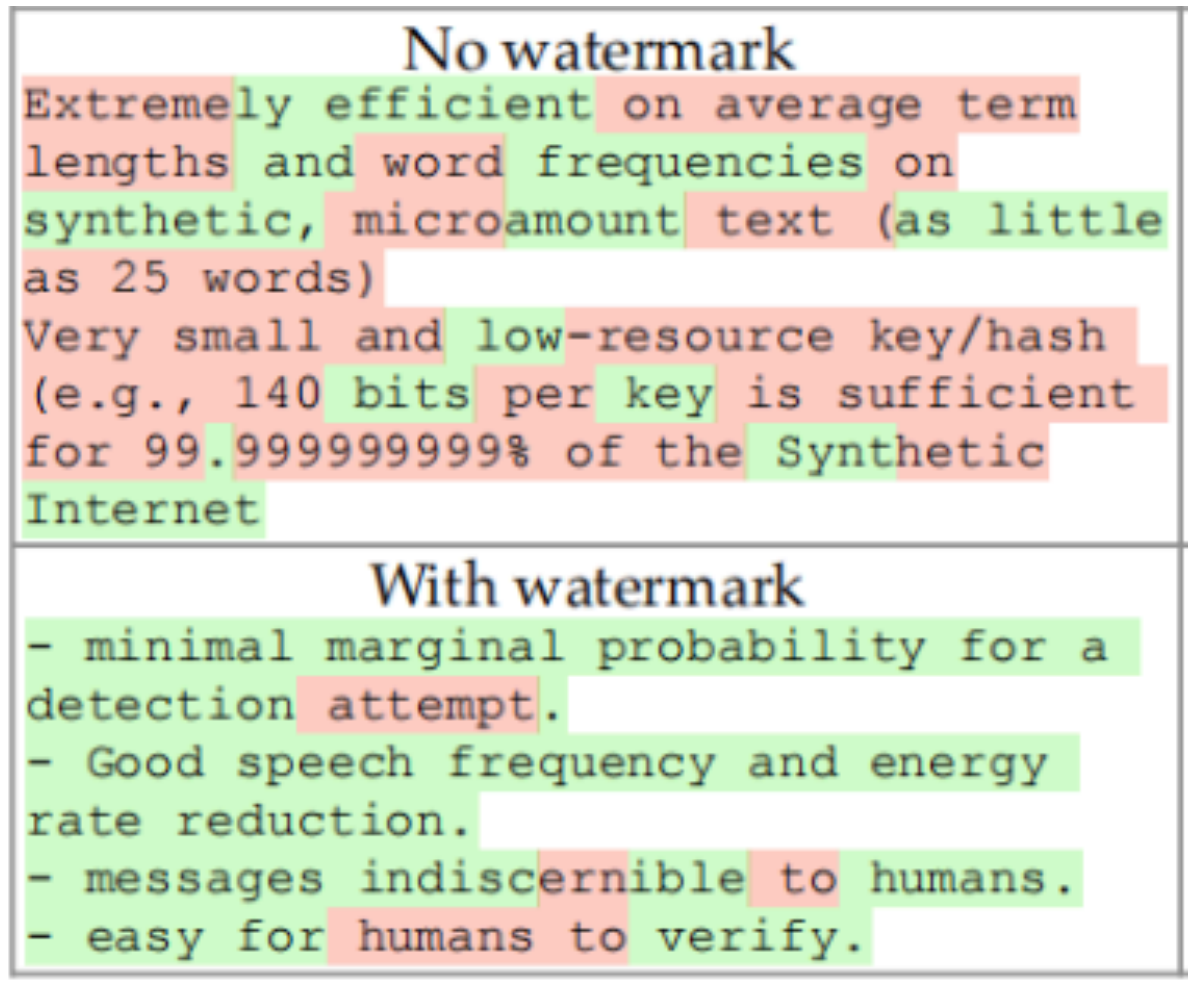
$$H_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Human/unwatermarked LLM ($Q_{X_1^T}$)

Watermarking scheme ($P_{\zeta_1^T}$, $P_{X_1^T, \zeta_1^T}$)

Existing watermarking methods: **heuristic**

Example
[KGW-1, 2023]



- High miss detection when requiring low false alarm

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

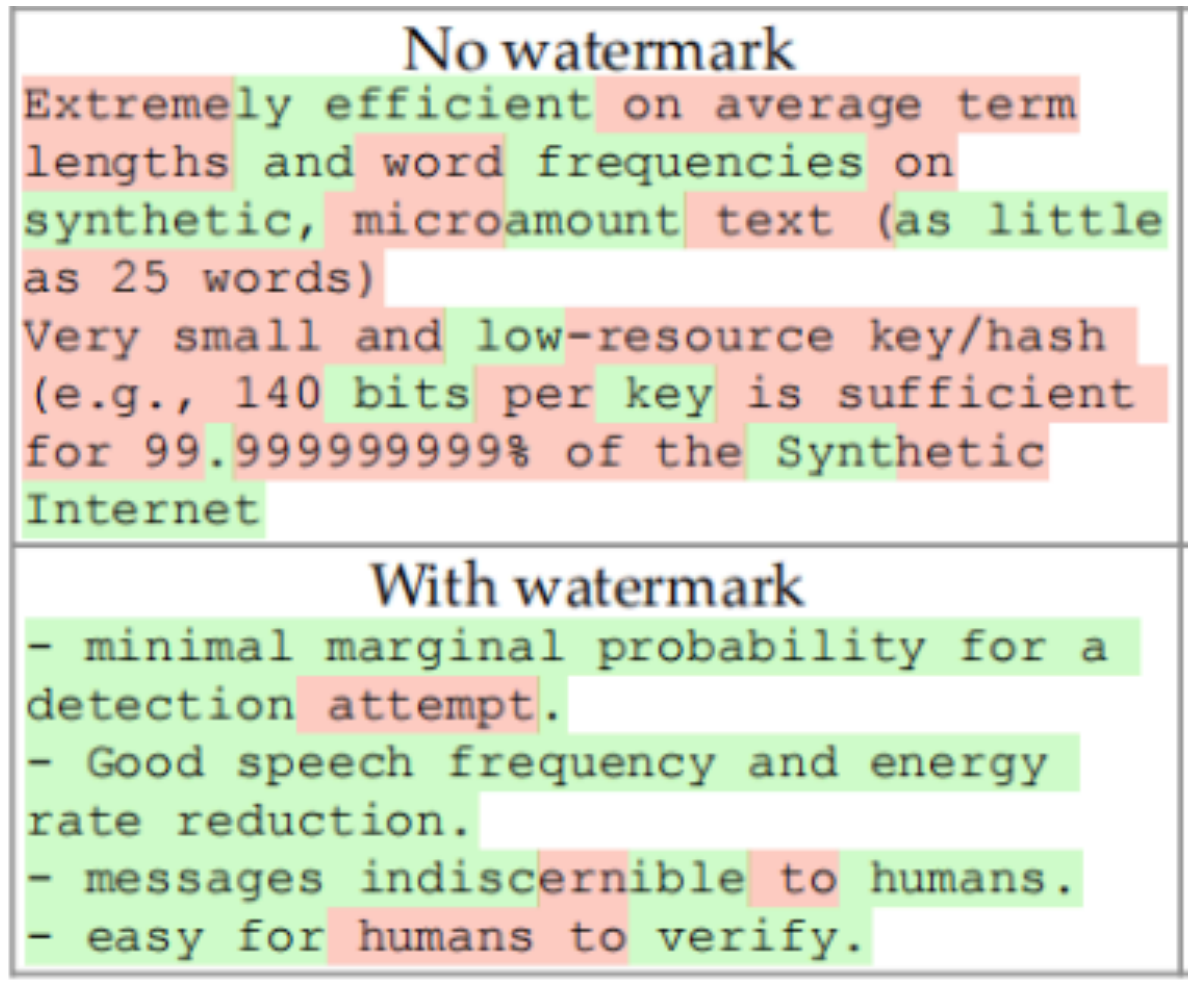
$$H_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$H_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Watermarking scheme

Existing watermarking methods: **heuristic**

Example
[KGW-1, 2023]



- High miss detection when requiring low false alarm
- Not distortion-free

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$$\mathrm{H}_0 : X_1^T \text{ is human written, i.e., } (X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$$

$$\mathrm{H}_1 : X_1^T \text{ is LLM generated, i.e., } (X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$$

Watermarking scheme

Find the best watermarking scheme & detector:

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$\mathrm{H}_0 : X_1^T$ is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

$\mathrm{H}_1 : X_1^T$ is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Find the best watermarking scheme & detector:

Minimize miss detection $\rightarrow$ $\min_{\gamma, P_{X_1^T, \zeta_1^T}} \beta_1(\gamma, P_{X_1^T, \zeta_1^T})$

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Find the best watermarking scheme & detector:

$$\min_{\gamma, P_{X_1^T, \zeta_1^T}} \beta_1(\gamma, P_{X_1^T, \zeta_1^T})$$

Humans are very creative, can write arbitrary texts

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$\mathrm{H}_0 : X_1^T$ is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

$\mathrm{H}_1 : X_1^T$ is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Find the best watermarking scheme & detector:

Humans are very creative, can write arbitrary texts

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \beta_1(\gamma, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$$

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Find the best watermarking scheme & detector:

$$\min_{\gamma,\, P_{X_1^T, \zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T, \zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$$

Ensure text quality

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Find the best watermarking scheme & detector:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$$

Ensure text quality $\rightarrow$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$$

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X_1^T is human written, i.e., $(X_1^T, \zeta_1^T) \sim Q_{X_1^T} \otimes P_{\zeta_1^T}$

H_1 : X_1^T is LLM generated, i.e., $(X_1^T, \zeta_1^T) \sim P_{X_1^T, \zeta_1^T}$

Watermarking scheme

Find the best watermarking scheme & detector:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$$

Fundamental Limit for Type-II Error

Optimization problem:

$$\min_{\gamma,\ P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma,\ P_{X_1^T,\zeta_1^T})$$

s.t. $$\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$$

$$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

Fundamental Limit for Type-II Error

Watermarked text distribution: $P^*_{X_1^T} = \arg \min_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma, P_{X_1^T, \zeta_1^T}} \beta_1(\gamma, P_{X_1^T, \zeta_1^T})$$

s.t. $\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$

$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$

✦ **Minimum Type-II error:**

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon) = \sum_{x_1^T} \left(P^*_{X_1^T}(x_1^T) - \alpha\right)_+$$

Fundamental Limit for Type-II Error

Watermarked text distribution: $P^*_{X_1^T} = \arg\min_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma, P_{X_1^T, \zeta_1^T}} \beta_1(\gamma, P_{X_1^T, \zeta_1^T})$$

s.t. $\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$

$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$

✦ **Minimum Type-II error:**

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon) = \sum_{x_1^T} \left(P^*_{X_1^T}(x_1^T) - \alpha\right)_+$$

Best achievable for any watermarking methods

Same as Huang et al. (2023, Theorem 3.2) but under different framework

Fundamental Limit for Type-II Error

Watermarked text distribution: $P^*_{X_1^T} = \arg \min_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma, P_{X_1^T, \zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T, \zeta_1^T})$$

s.t. $\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$

$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$

✦ **Minimum Type-II error:**

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon) = \sum_{x_1^T} \left(P^*_{X_1^T}(x_1^T) - \alpha \right)_+$$

Fundamental Limit for Type-II Error

Watermarked text distribution: $P^*_{X_1^T} = \arg\min_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma, P_{X_1^T, \zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T, \zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$$

$$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$$

D_{TV}

✦ Minimum Type-II error:

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon) = \sum_{x_1^T} \left(P^*_{X_1^T}(x_1^T) - \alpha\right)_+$$

Fundamental Limit for Type-II Error

Watermarked text distribution: $P^*_{X_1^T} = \arg \min_{P_{X_1^T}: \mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T, \zeta_1^T}} \quad \beta_1(\gamma,\, P_{X_1^T, \zeta_1^T})$$

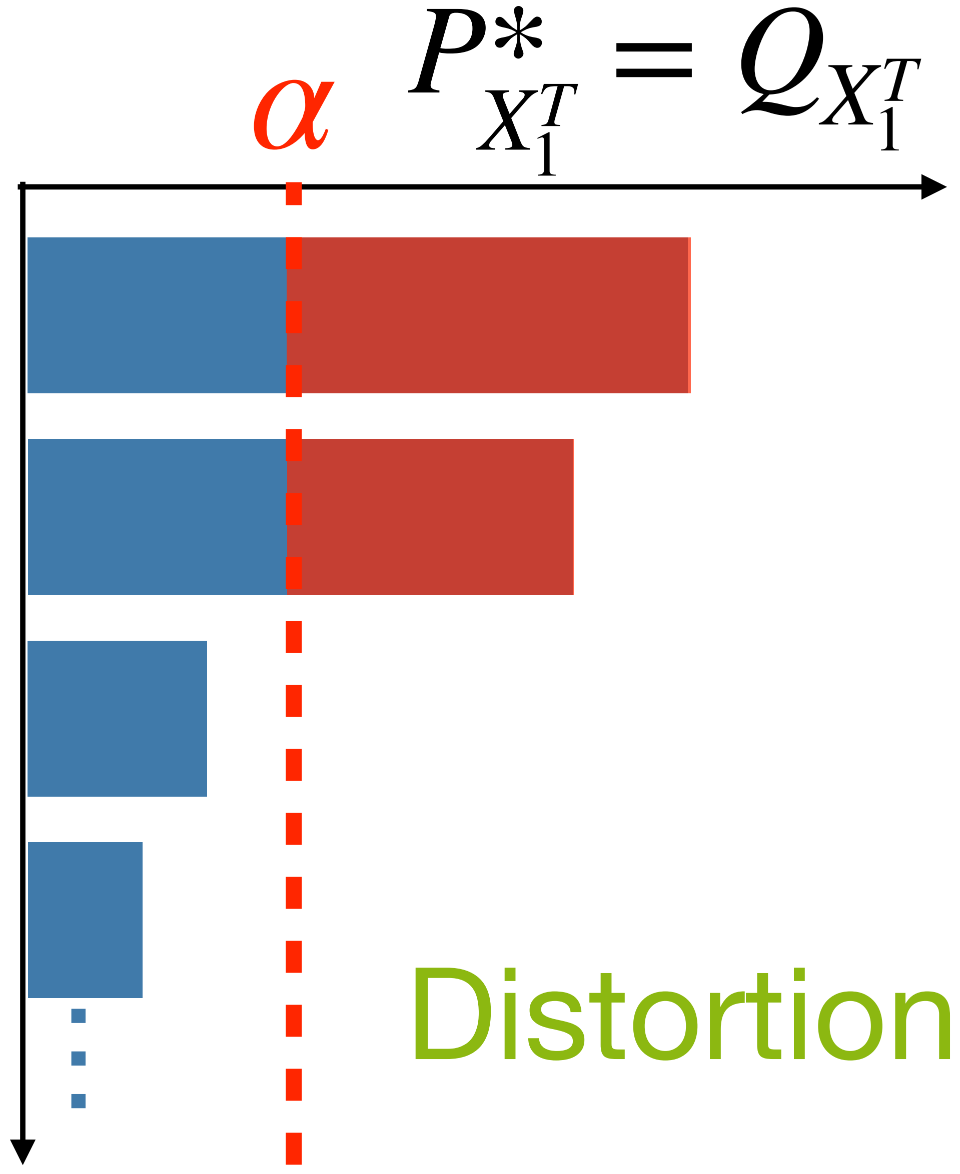
$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

D_{TV}

✦ **Minimum Type-II error:**

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon) = \sum_{x_1^T} \left(P^*_{X_1^T}(x_1^T) - \alpha\right)_+$$



Distortion-free

Fundamental Limit for Type-II Error

Watermarked text distribution: $P^*_{X_1^T} = \arg \min_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma, P_{X_1^T, \zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T, \zeta_1^T})$$

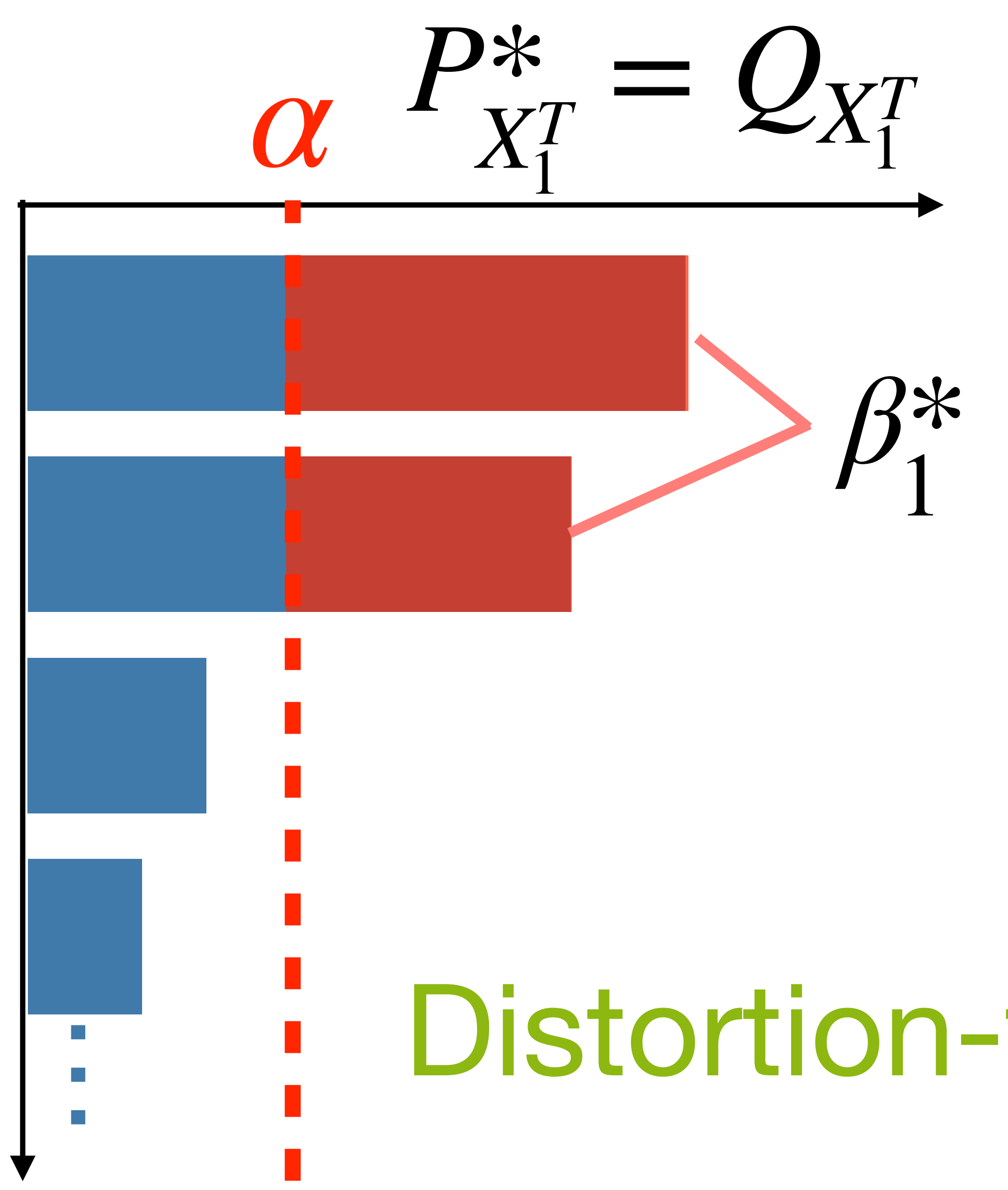
$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$$

$$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

D_{TV}

✦ **Minimum Type-II error:**

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon) = \sum_{x_1^T} \left(P^*_{X_1^T}(x_1^T) - \alpha\right)_+$$



Distortion-free

Fundamental Limit for Type-II Error

Watermarked text distribution: $P^*_{X_1^T} = \arg\min_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma, P_{X_1^T,\zeta_1^T}} \beta_1(\gamma, P_{X_1^T,\zeta_1^T})$$

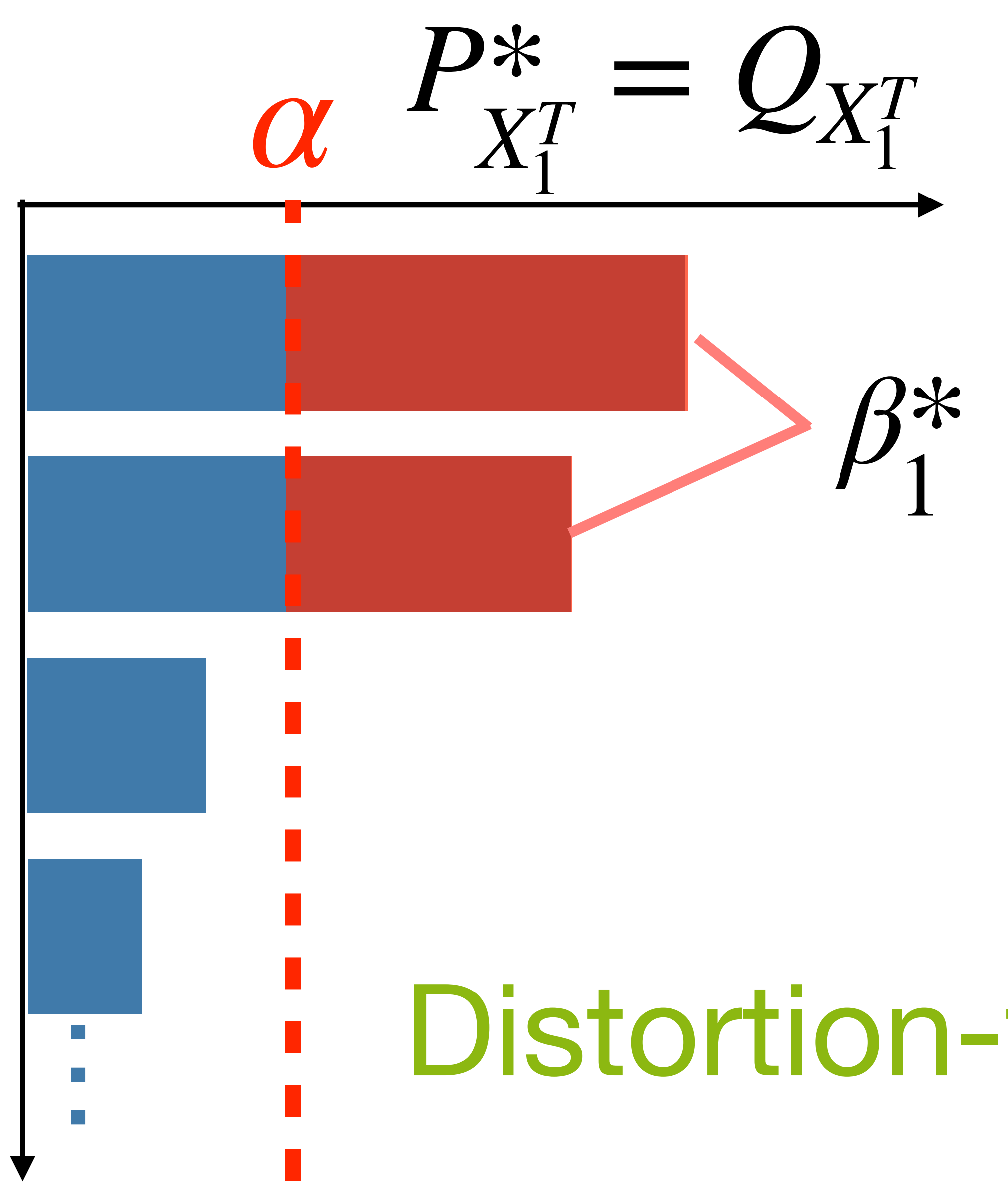
s.t. $\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$

$$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

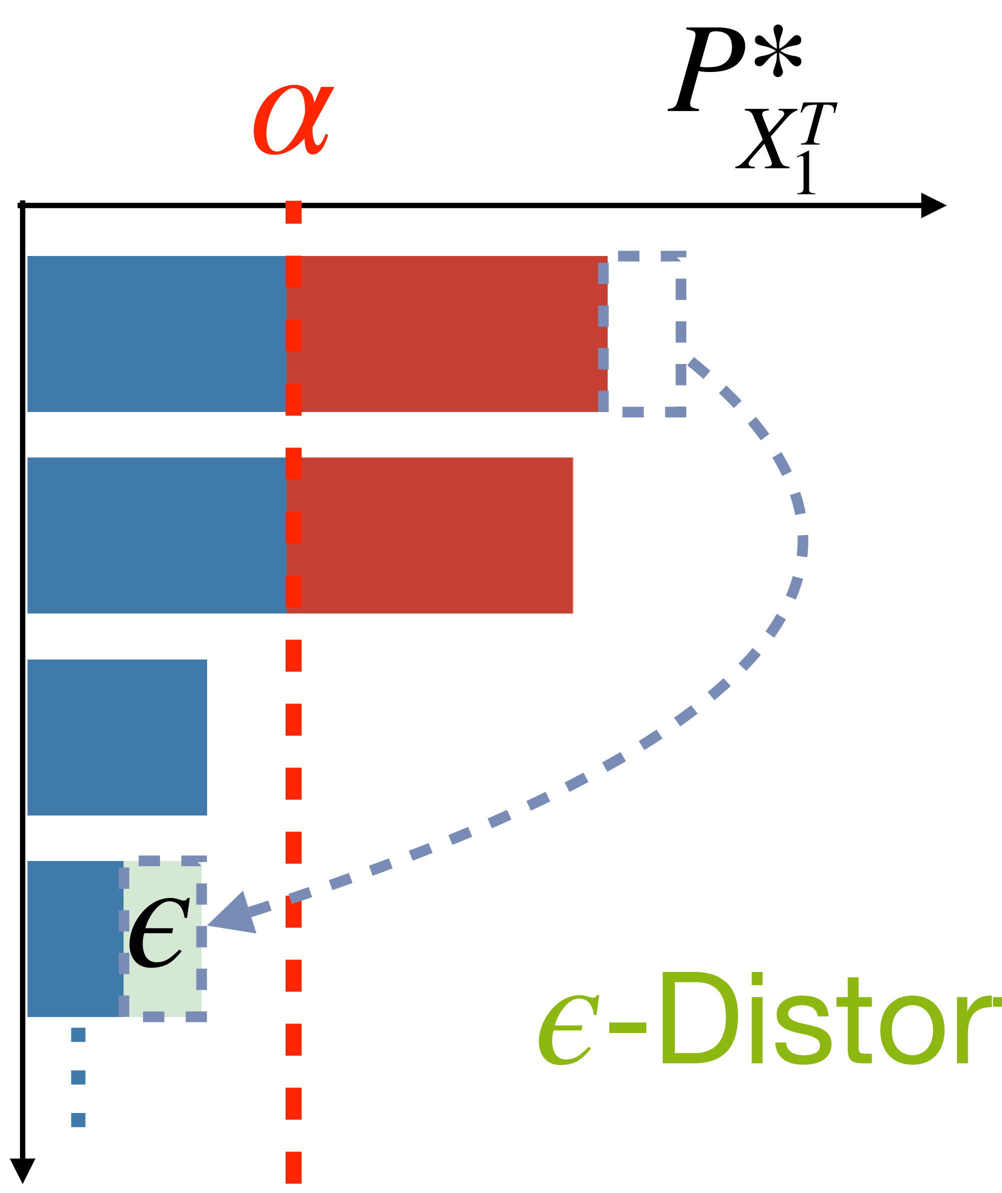
D_{TV}

✦ **Minimum Type-II error:**

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon) = \sum_{x_1^T} \left(P^*_{X_1^T}(x_1^T) - \alpha\right)_+$$



Distortion-free



ϵ-Distorted

Fundamental Limit for Type-II Error

Watermarked text distribution: $P^*_{X_1^T} = \arg\min_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \beta_1(\gamma, P_{X_1^T,\zeta_1^T})$$

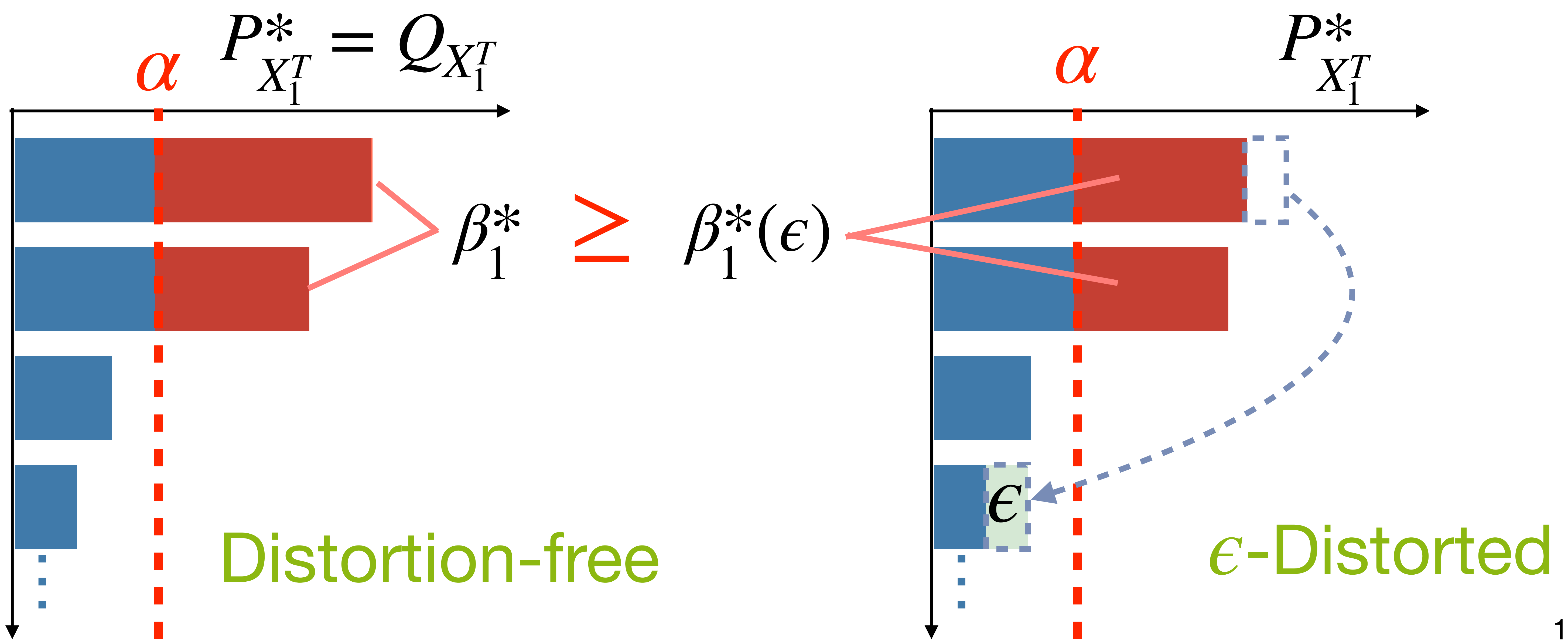
s.t. $\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$

$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$

D_{TV}

✦ **Minimum Type-II error:**

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon) = \sum_{x_1^T} \left(P^*_{X_1^T}(x_1^T) - \alpha\right)_+$$



Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma,\, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$$

Jointly Optimal Detector and Watermarking Scheme

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma,\, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$$

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma,\, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g : \mathscr{Z}^T \to \mathscr{S} \supset \mathscr{V}^T$

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma,\, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$$

$$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha \qquad (\Delta_1)$$

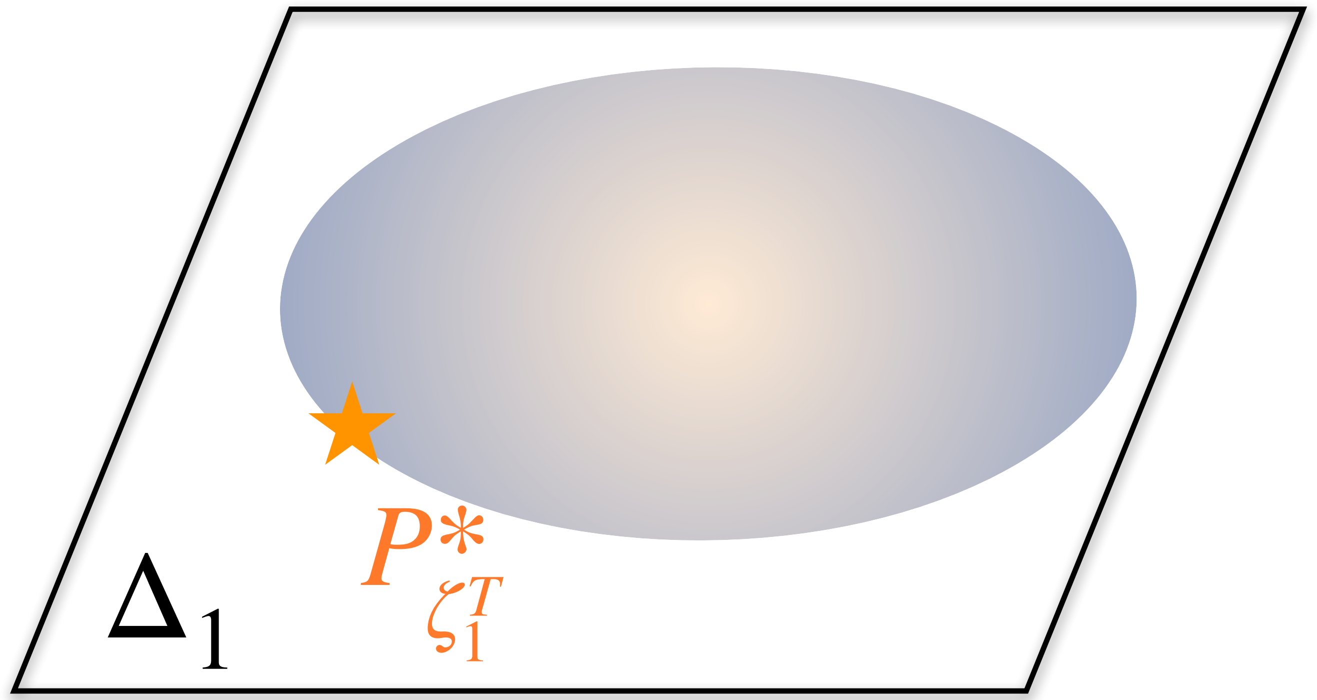
$$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:



Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma,\, P_{X_1^T,\zeta_1^T})$$

s.t. $\sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$ $\quad(\Delta_1)$

$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$ $\quad(\Delta_2)$

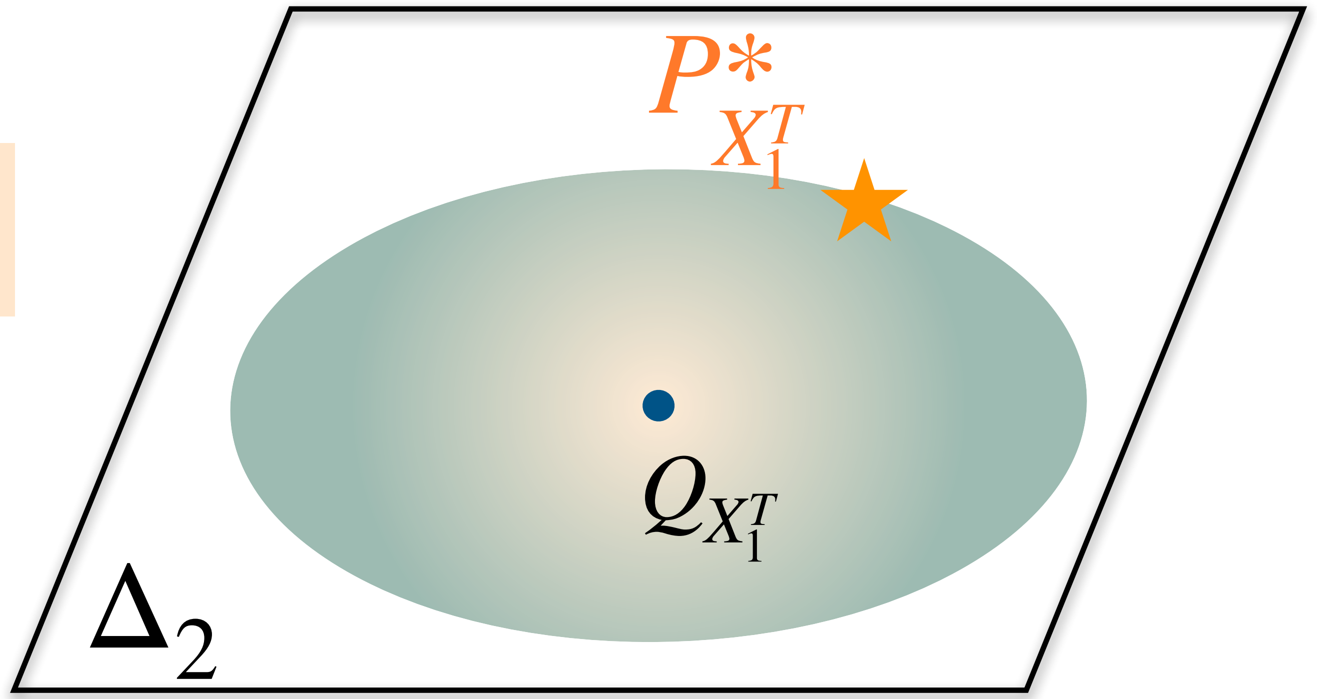
$$P^*_{X_1^T} = \arg\min_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$$

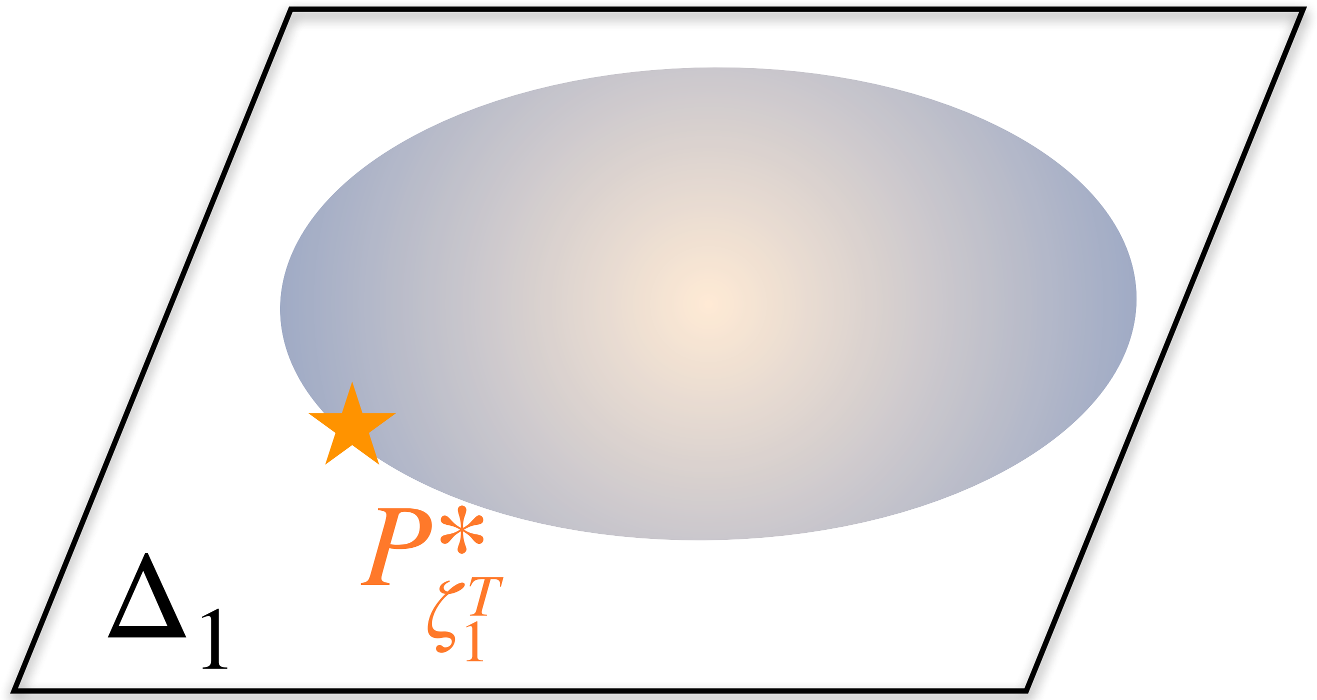
✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:





Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma,\, P_{X_1^T,\zeta_1^T}) = \mathbb{E}_{P_{X_1^T,\zeta_1^T}}[1-\gamma(X_1^T,\zeta_1^T)]$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha \qquad (\Delta_1)$$

$$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon \qquad (\Delta_2)$$

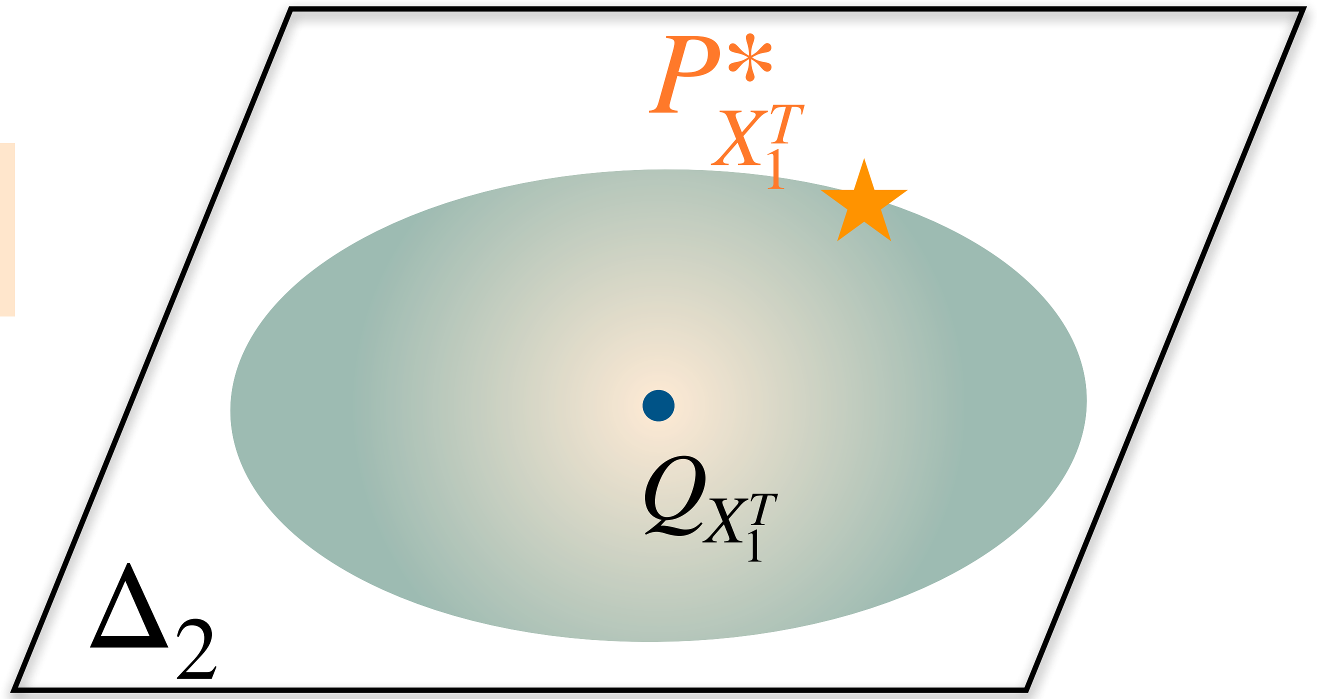
$$P^*_{X_1^T} = \arg \min_{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\le\epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T)-\alpha)_+$$

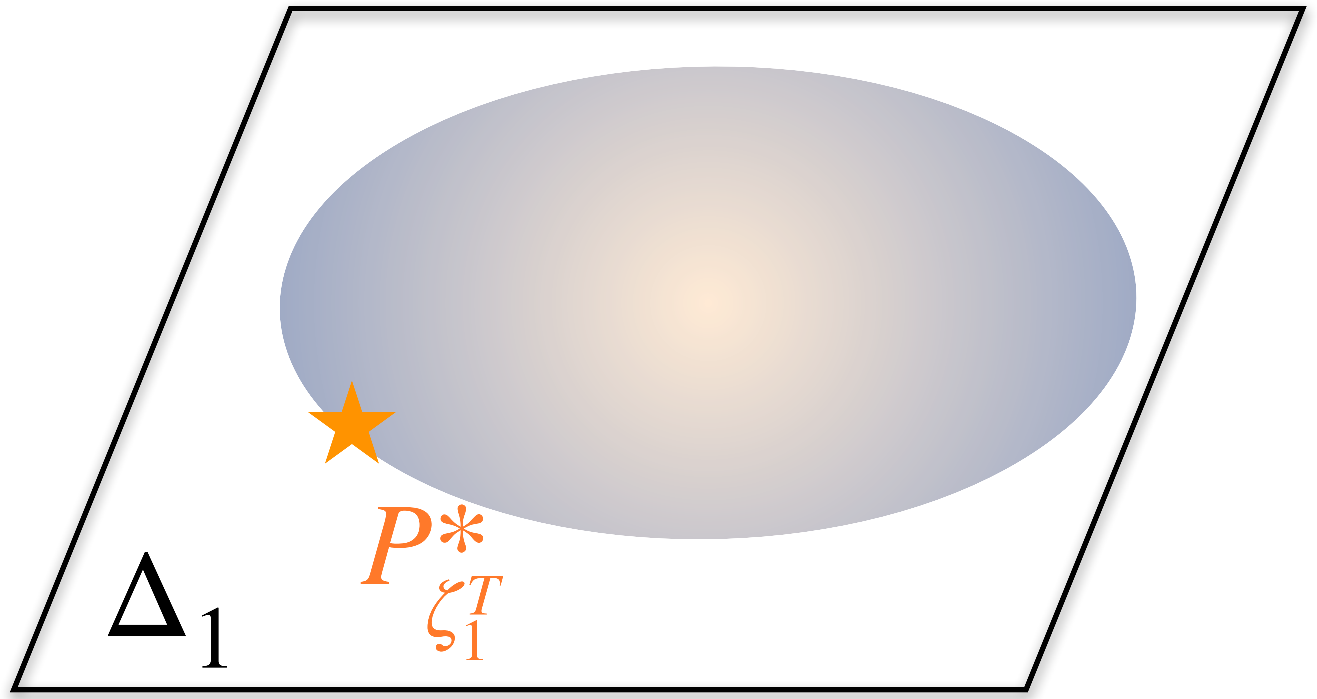
✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g: \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:





Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \beta_1(\gamma, P_{X_1^T,\zeta_1^T}) = \mathbb{E}_{P_{X_1^T,\zeta_1^T}}[1-\gamma(X_1^T,\zeta_1^T)]$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha \qquad (\Delta_1)$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon \qquad (\Delta_2)$$

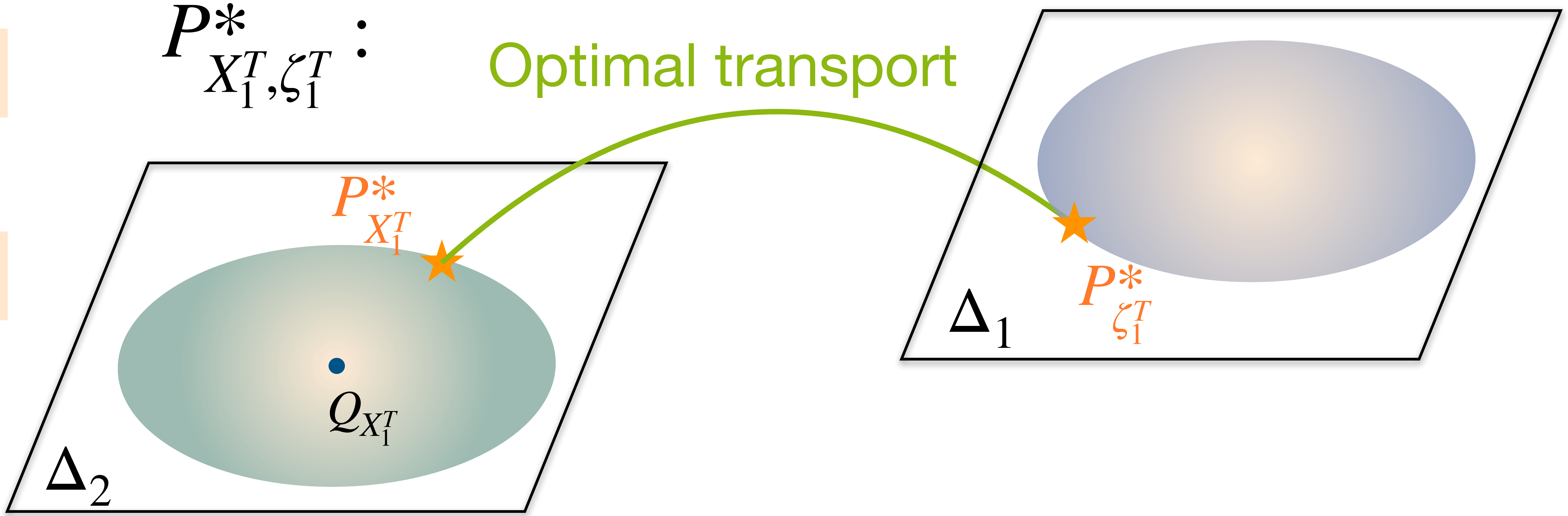
$$P^*_{X_1^T} = \arg\min_{P_{X_1^T}:\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:



Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T,\zeta_1^T}) = \mathbb{E}_{P_{X_1^T,\zeta_1^T}}[1-\gamma(X_1^T,\zeta_1^T)]$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha \qquad (\Delta_1)$$

$$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon \qquad (\Delta_2)$$

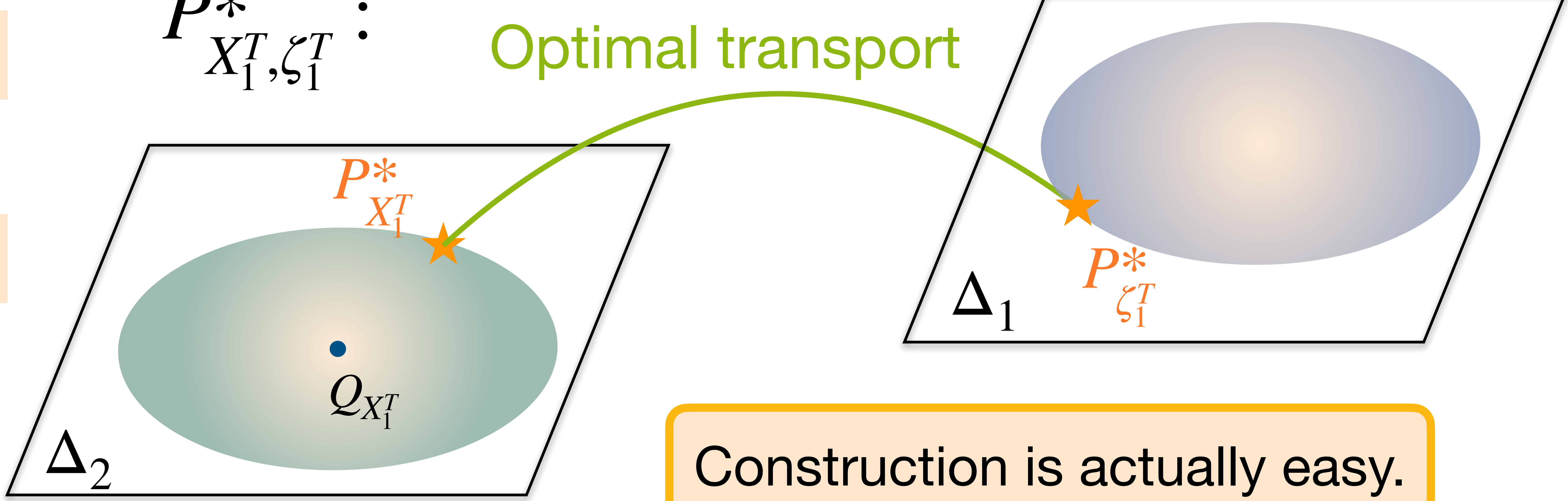
$$P^*_{X_1^T} = \arg\min_{P_{X_1^T}:\mathsf{D}(P_{X_1^T},Q_{X_1^T})\le\epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T)-\alpha)_+$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:



Construction is actually easy.

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g : \mathcal{X}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:

$$P^*_{X_1^T} = \arg \min_{P_{X_1^T}: \mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$$

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma,\, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

✦ **Jointly optimal detector γ^***

and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g : \mathcal{X}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:

$(T = 1)$

$$P^*_{X_1^T} = \arg \min_{P_{X_1^T} : \mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$$

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma,\, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$$

$$\mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

$$P^*_{X_1^T} = \arg \min_{P_{X_1^T}: \mathsf{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon} \sum_{x_1^T} (P_{X_1^T}(x_1^T) - \alpha)_+$$

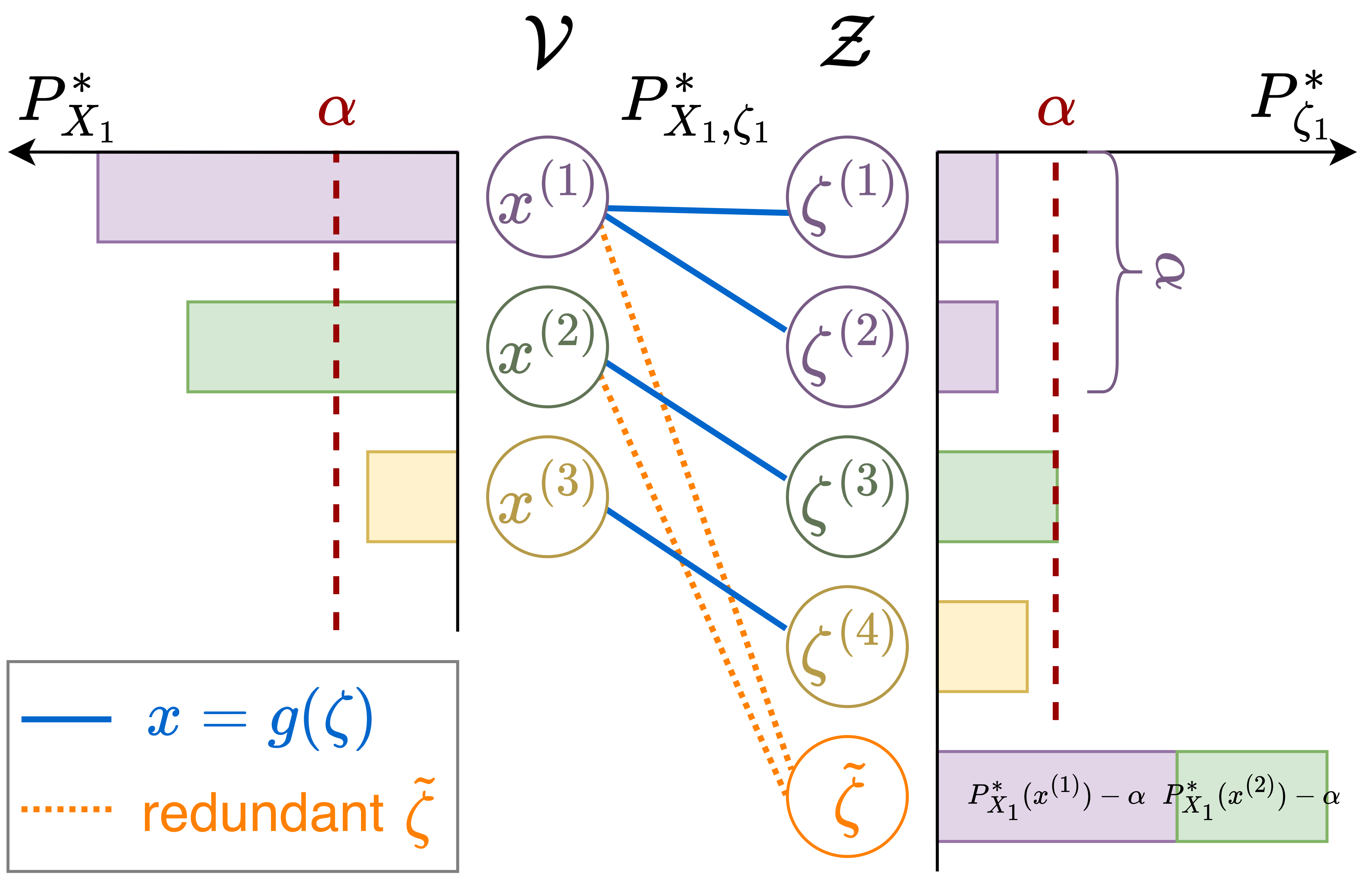
✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g: \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:

$(T = 1)$



Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \le \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:**

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:

$P^*_{\zeta_1^T}$ **adaptive** to original LLM predicted distribution $Q_{X_1^T}$

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma, P_{X_1^T,\zeta_1^T}} \beta_1(\gamma, P_{X_1^T,\zeta_1^T})$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}) \leq \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$$

✦ **Jointly optimal detector γ^***
and watermarking scheme $P^*_{X_1^T,\zeta_1^T}$:

$$\gamma^* = \mathbf{1}\{X_1^T = g(\zeta_1^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X_1^T,\zeta_1^T}$:

$P^*_{\zeta_1^T}$ **adaptive** to original LLM predicted distribution $Q_{X_1^T}$

↓

Unlike existing watermarking methods

Sequence-Level Optimal to Token-Level Optimal

Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for fixed $T \Rightarrow$ unable to implement dynamically

Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for fixed $T \Rightarrow$ unable to implement dynamically
- Solution: implement it token by token

Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for fixed $T \Rightarrow$ unable to implement dynamically
- Solution: implement it token by token

Detector:

Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for fixed T $\Rightarrow$ unable to implement dynamically
- Solution: implement it token by token

Detector:

$$\gamma_{tk} = \mathbf{1}\left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{X_t = g(\zeta_t)\} \geq \lambda \right\}$$ for some surjective $g : \mathscr{Z} \to \mathscr{S} \supset \mathscr{V}$

Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for fixed $T \Rightarrow$ unable to implement dynamically
- Solution: implement it token by token

Detector:

$$\gamma_{tk} = \mathbf{1}\left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{X_t = g(\zeta_t)\} \geq \lambda \right\}$$ for some surjective $g : \mathcal{Z} \to \mathcal{S} \supset \mathcal{V}$

Watermarking scheme:

$$P^*_{X_1^T, \zeta_1^T} \to P^*_{X_t, \zeta_t | X_1^{t-1}, \zeta_1^{t-1}}, \forall t = 1, 2, \ldots, T$$

Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for fixed $T \Rightarrow$ unable to implement dynamically
- Solution: implement it token by token

Detector:

$$\gamma_{tk} = \mathbf{1}\left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{X_t = g(\zeta_t)\} \geq \lambda \right\}$$ for some surjective $g : \mathcal{X} \to \mathcal{S} \supset \mathcal{V}$

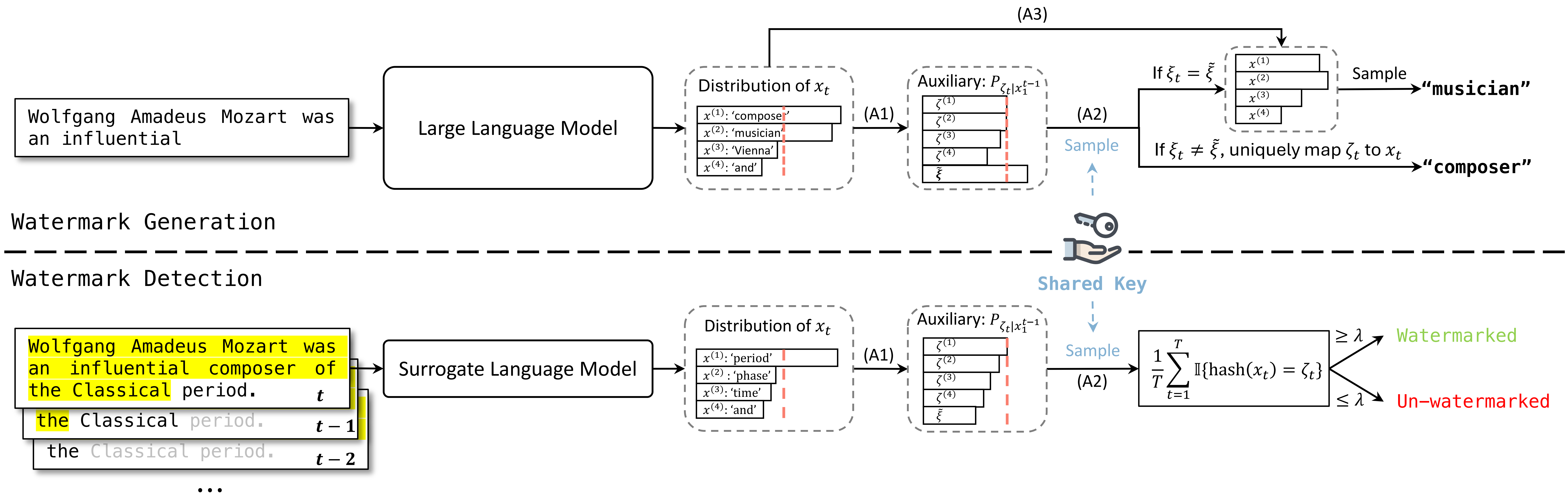
Watermarking scheme:

$$P^*_{X_1^T, \zeta_1^T} \to P^*_{X_t, \zeta_t | X_1^{t-1}, \zeta_1^{t-1}}, \forall t = 1, 2, \ldots, T$$

dependent

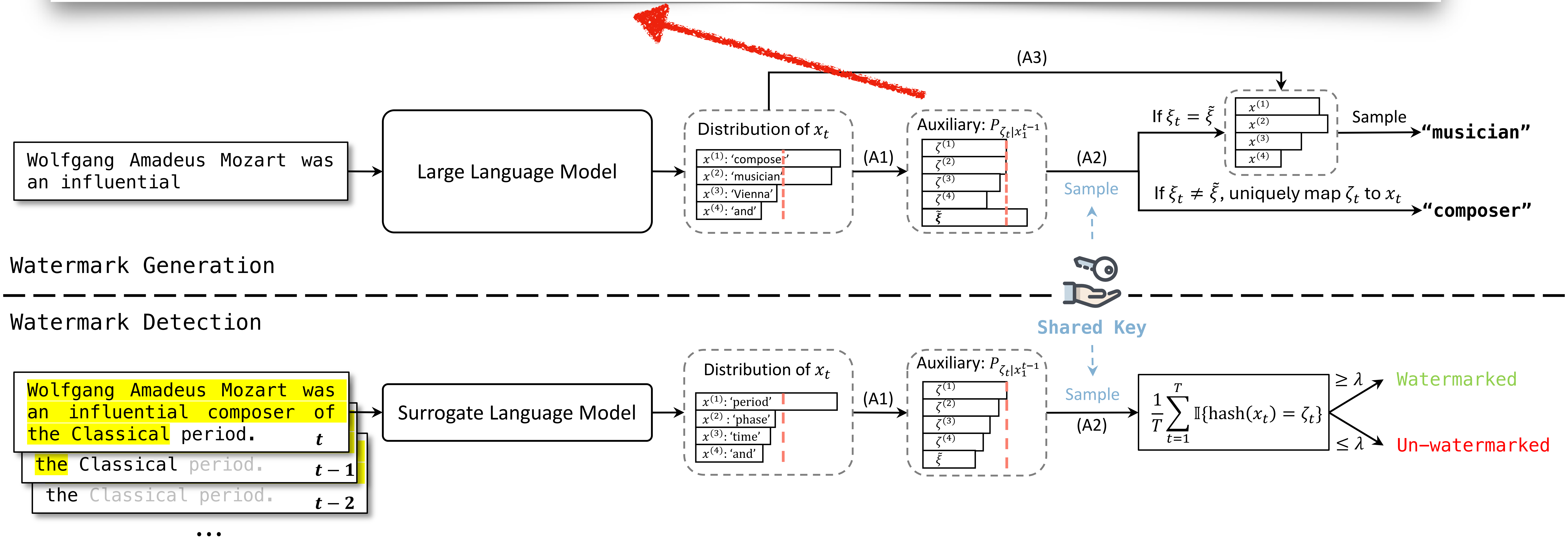
Token-level false alarm rate η $\xrightarrow{\text{controls}}$ Sequence-level false alarm α

DAWA: Distribution-Adaptive Watermarking Algorithm ($\epsilon = 0$, distortion-free)



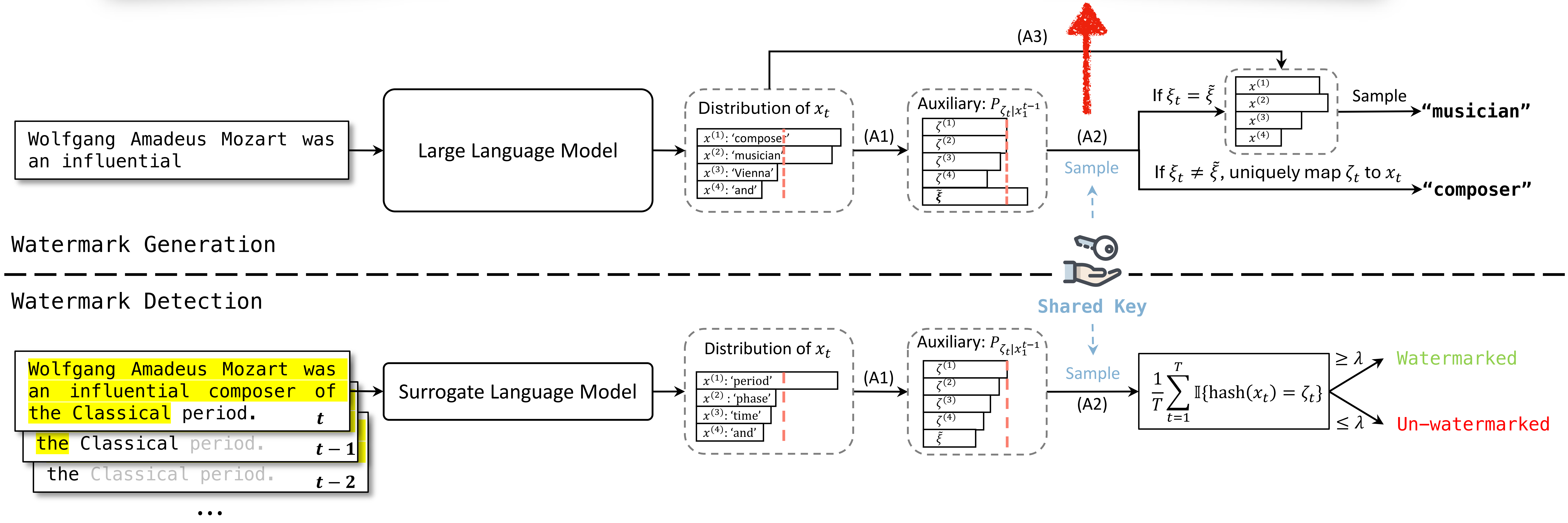
DAWA: Distribution-Adaptive Watermarking Algorithm ($\epsilon = 0$, distortion-free)

At each time t, construct $P^*_{\zeta_t|X_1^t}$ from the LLM predicted distribution $Q_{X_t|X_1^{t-1}}$

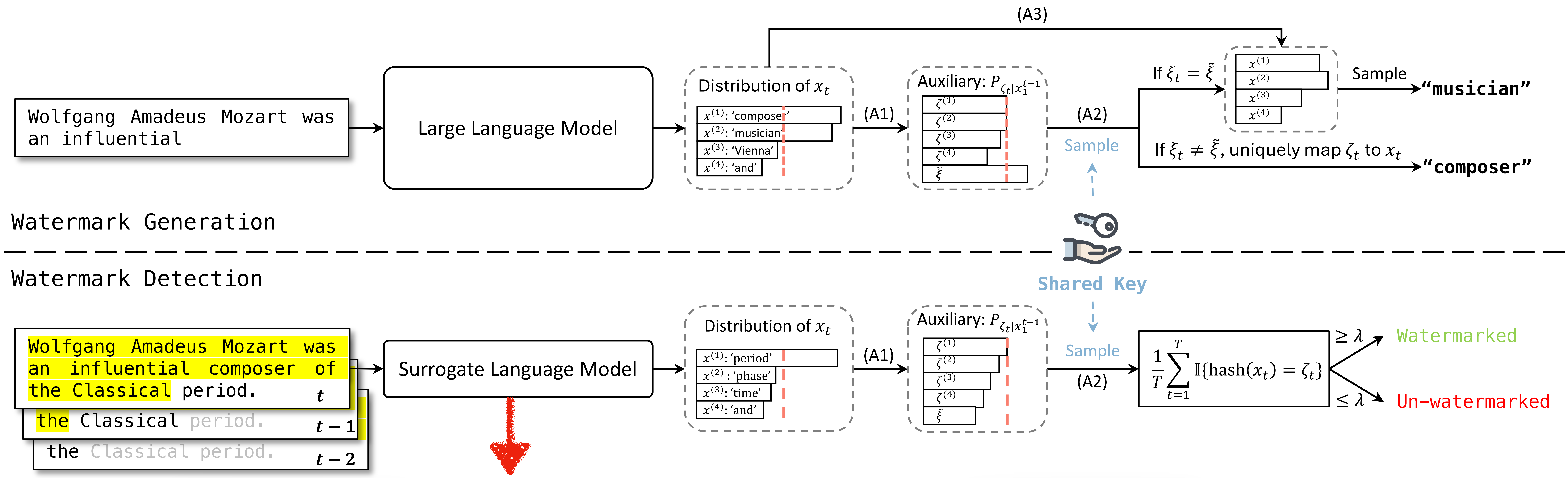


DAWA: Distribution-Adaptive Watermarking Algorithm ($\epsilon = 0$, distortion-free)

Sample ζ_t using Gumbel max trick: $\zeta_t \leftarrow \arg\max_{\zeta} \log P^*_{\zeta_t | x_1^t}(\zeta) + G_{\zeta,t}$

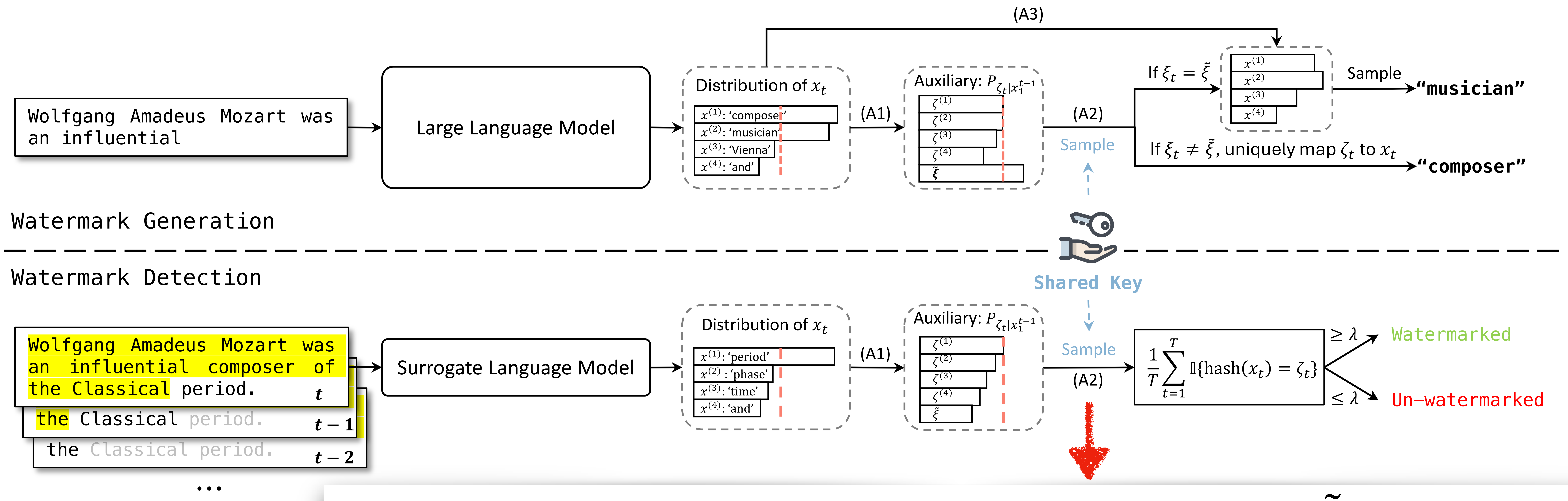


DAWA: Distribution-Adaptive Watermarking Algorithm ($\epsilon = 0$, distortion-free)



Approximate distribution of X_t so as to construct $\tilde{P}_{\zeta_t|x_1^t}$

DAWA: Distribution-Adaptive Watermarking Algorithm ($\epsilon = 0$, distortion-free)



Sample ζ_t using Gumbel max trick: $\zeta_t \leftarrow \arg\max_{\zeta} \log \tilde{P}_{\zeta_t|x_1^t}(\zeta) + G_{\zeta,t}$

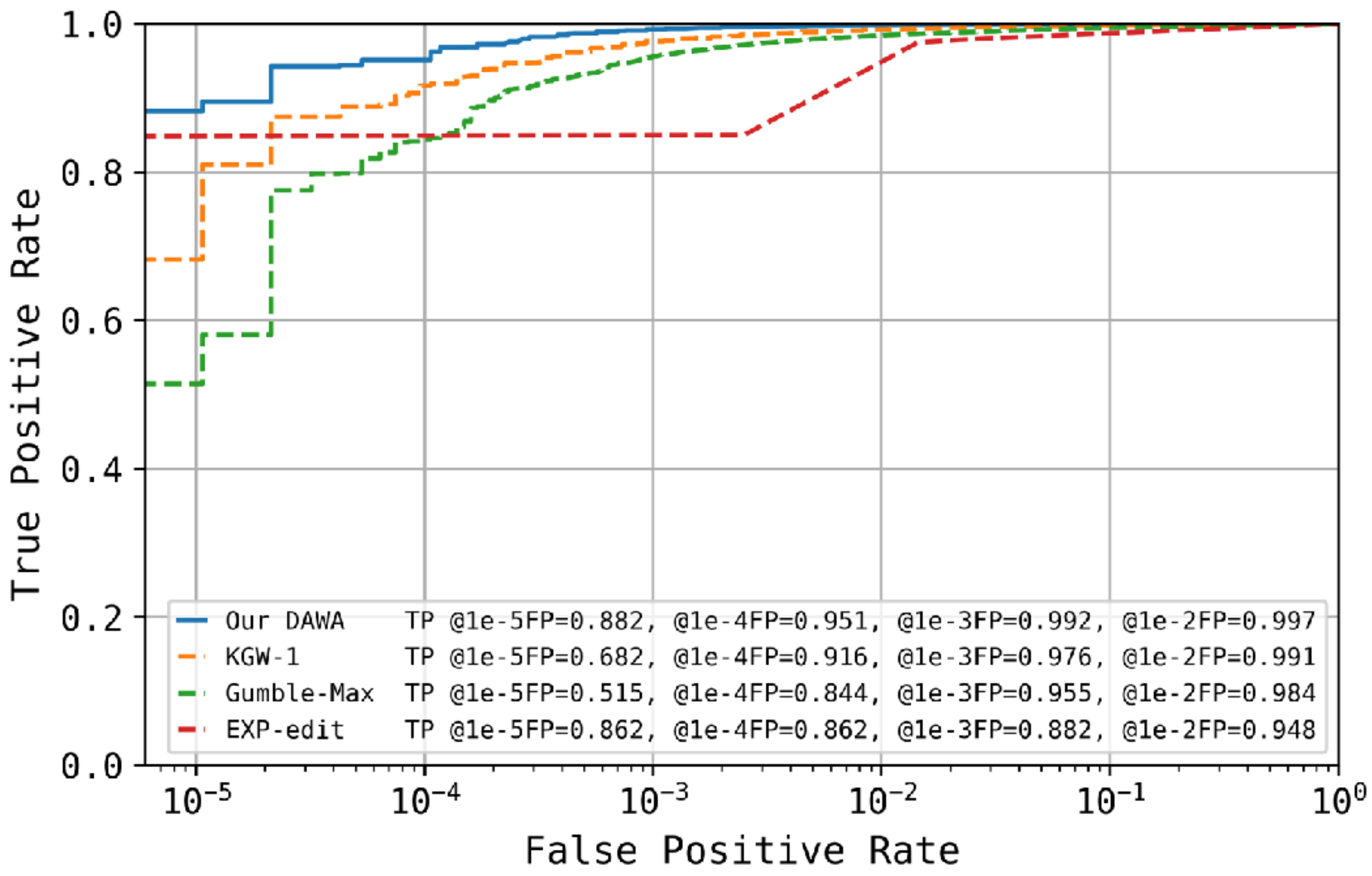
Experimental Result

DAWA (**D**istribution-**A**daptive **W**atermarking **A**lgorithm)

Experimental Result

DAWA (**D**istribution-**A**daptive **W**atermarking **A**lgorithm)

Fast and Accurate



Experimental Result

DAWA (**D**istribution-**A**daptive **W**atermarking **A**lgorithm)

Fast and Accurate

Text quality high

Methods	Human	KGW-1	EXP-Edit	Gumbel-Max	**Ours**
BLEU Score	0.219	0.158	0.203	0.210	0.214
Avg Perplexity	8.846	14.327	12.186	11.732	6.495

Future Work: With Text Modifications?

Future Work: With Text Modifications?

- $f : \mathscr{V}^T \to [K]$: a function that maps a sequence of tokens X_1^T to a finite latent space $[K]$, e.g., a semantic mapping

Future Work: With Text Modifications?

- $f : \mathscr{V}^T \to [K]$: a function that maps a sequence of tokens X_1^T to a finite latent space $[K]$, e.g., a semantic mapping
- $\mathscr{B}_f(x_1^T) = \{\tilde{x}_1^T \in \mathscr{V}^T : f(\tilde{x}_1^T) = f(x_1^T)\}$ be an equivalence class containing x_1^T

Future Work: With Text Modifications?

- $f : \mathscr{V}^T \to [K]$: a function that maps a sequence of tokens X_1^T to a finite latent space $[K]$, e.g., a semantic mapping
- $\mathscr{B}_f(x_1^T) = \{\tilde{x}_1^T \in \mathscr{V}^T : f(\tilde{x}_1^T) = f(x_1^T)\}$ be an equivalence class containing x_1^T
- Text modification: x_1^T can be modified as any text within $\mathscr{B}_f(x_1^T)$

Future Work: With Text Modifications?

- $f : \mathscr{V}^T \to [K]$: a function that maps a sequence of tokens X_1^T to a finite latent space $[K]$, e.g., a semantic mapping
- $\mathscr{B}_f(x_1^T) = \{\tilde{x}_1^T \in \mathscr{V}^T : f(\tilde{x}_1^T) = f(x_1^T)\}$ be an equivalence class containing x_1^T
- Text modification: x_1^T can be modified as any text within $\mathscr{B}_f(x_1^T)$
- f-robust Type-I and Type-II errors:

Future Work: With Text Modifications?

- $f : \mathscr{V}^T \to [K]$: a function that maps a sequence of tokens X_1^T to a finite latent space $[K]$, e.g., a semantic mapping
- $\mathscr{B}_f(x_1^T) = \{\tilde{x}_1^T \in \mathscr{V}^T : f(\tilde{x}_1^T) = f(x_1^T)\}$ be an equivalence class containing x_1^T
- Text modification: x_1^T can be modified as any text within $\mathscr{B}_f(x_1^T)$
- f-robust Type-I and Type-II errors:

$$\beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}, f) := \mathbb{E}_{Q_{X_1^T} \otimes P_{\zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathscr{B}_f(X_1^T)} \mathbf{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \right]$$

Future Work: With Text Modifications?

- $f : \mathscr{V}^T \to [K]$: a function that maps a sequence of tokens X_1^T to a finite latent space $[K]$, e.g., a semantic mapping
- $\mathscr{B}_f(x_1^T) = \{\tilde{x}_1^T \in \mathscr{V}^T : f(\tilde{x}_1^T) = f(x_1^T)\}$ be an equivalence class containing x_1^T
- Text modification: x_1^T can be modified as any text within $\mathscr{B}_f(x_1^T)$
- f-robust Type-I and Type-II errors:

$$\beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}, f) := \mathbb{E}_{Q_{X_1^T} \otimes P_{\zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathscr{B}_f(X_1^T)} \mathbf{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 1\} \right]$$

$$\beta_1(\gamma, P_{X_1^T, \zeta_1^T}, f) := \mathbb{E}_{P_{X_1^T, \zeta_1^T}} \left[\sup_{\tilde{x}_1^T \in \mathscr{B}_f(X_1^T)} \mathbf{1}\{\gamma(\tilde{x}_1^T, \zeta_1^T) = 0\} \right]$$

Robustness Against Text Modifications

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma,\, P_{X_1^T,\zeta_1^T},\, f)$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma,\, Q_{X_1^T},\, P_{\zeta_1^T},\, f) \leq \alpha$$

$$\mathrm{D}(P_{X_1^T},\, Q_{X_1^T}) \leq \epsilon$$

Robustness Against Text Modifications

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T,\zeta_1^T}, f)$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}, f) \le \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

✦ **Minimum f-robust Type-II error:**

Robustness Against Text Modifications

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T,\zeta_1^T}, f)$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}, f) \le \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \le \epsilon$$

✦ **Minimum f-robust Type-II error:**

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon, f) = \min_{P_{X_1^T}:\mathrm{D}(P_{X_1^T},Q_{X_1^T})\le\epsilon} \sum_{k\in[K]} \left(\left(\sum_{x_1^T: f(x_1^T)=k} P_{X_1^T}(x_1^T) \right) - \alpha \right)_+$$

Robustness Against Text Modifications

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T,\zeta_1^T}, f)$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}, f) \leq \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$$

✦ **Minimum f-robust Type-II error:**

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon, f)$$

$$= \min_{P_{X_1^T}:\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon} \sum_{k\in[K]} \left(\left(\sum_{x_1^T: f(x_1^T)=k} P_{X_1^T}(x_1^T) \right) - \alpha \right)_+$$

Higher than the minimum Type-II error without considering robustness

Robustness Against Text Modifications

Optimization problem:

$$\min_{\gamma,\, P_{X_1^T,\zeta_1^T}} \quad \beta_1(\gamma, P_{X_1^T,\zeta_1^T}, f)$$

$$\text{s.t.} \quad \sup_{Q_{X_1^T}} \beta_0(\gamma, Q_{X_1^T}, P_{\zeta_1^T}, f) \leq \alpha$$

$$\mathrm{D}(P_{X_1^T}, Q_{X_1^T}) \leq \epsilon$$

✦ **Minimum f-robust Type-II error:**

$$\beta_1^*(Q_{X_1^T}, \alpha, \epsilon, f)$$

$$= \min_{P_{X_1^T}:\mathrm{D}(P_{X_1^T},Q_{X_1^T})\leq\epsilon} \sum_{k\in[K]} \left(\left(\sum_{x_1^T: f(x_1^T)=k} P_{X_1^T}(x_1^T) \right) - \alpha \right)_+$$

Higher than the minimum Type-II error without considering robustness

✦ **Optimal watermarking scheme:**

add signal ζ_1^T to $P_{f(X_1^T)}$, e.g., in the semantic space